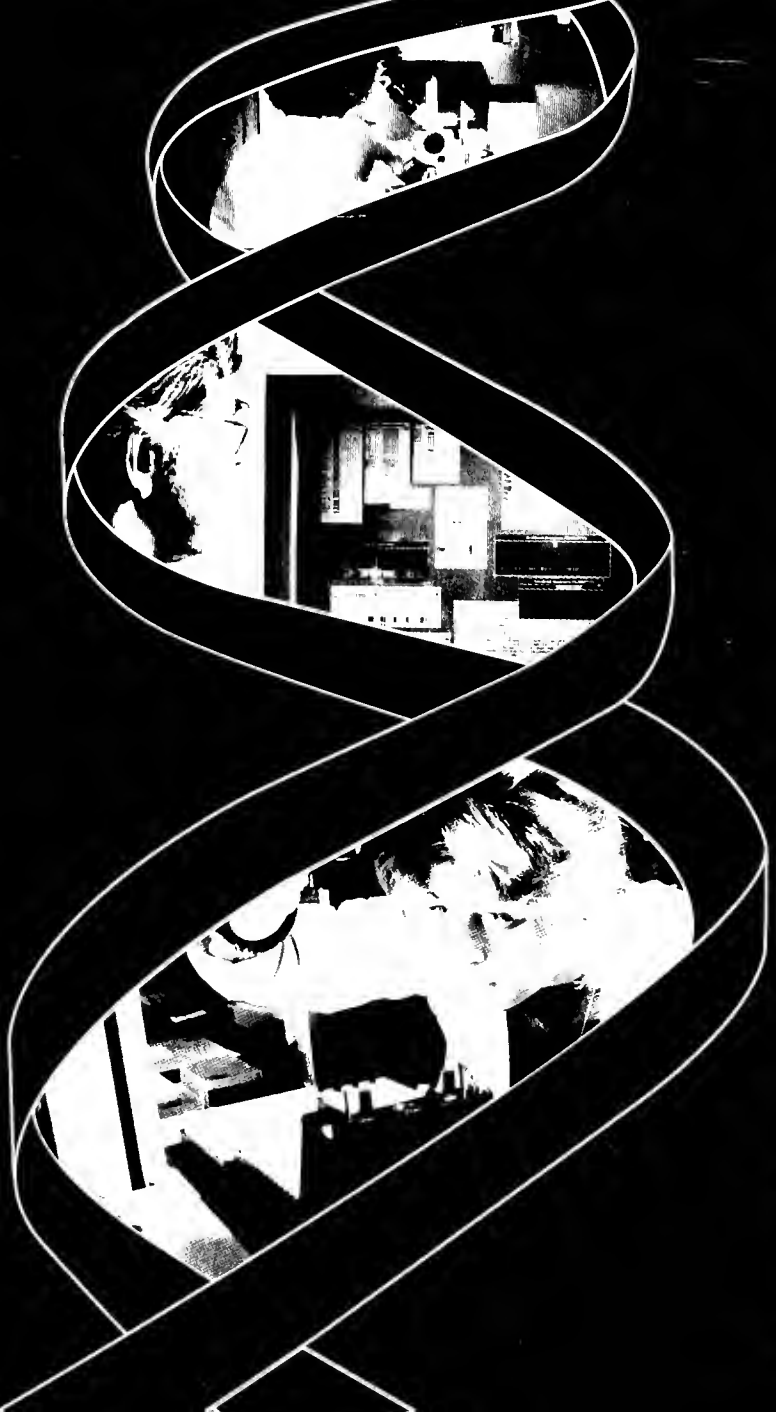
DOE/ER-0446P

Human Genome



1989–90 Program Report

United States Department of Energy

Office of Energy Research

Office of Health and Environmental Research

March 1990

Please address queries on this publication to:

Dr. Benjamin J. Barnhart
Manager, Human Genome Program
Office of Health and Environmental Research
U.S. Department of Energy
ER-72 GTN
Washington, DC 20585
(301) 353-5037, FTS 233-5037
FAX: (301) 353-5051, FTS 233-5051

Human Genome Management Information System
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6050
(615) 576-6669, FTS 626-6669
FAX: (615) 574-9888, FTS 624-9888

DOE/ER-0446P

Human Genome

1989–90 Program Report

Date Published: March 1990

U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research
Washington, D.C. 20585

Major Events in the Development of the DOE Human Genome Program

Santa Fe Meeting

Human Genome Initiative announced

Pilot projects pursued at national laboratories:

- Computer modeling and optimization of library ordering strategies;
- Database management support;
- Advanced instrumentation and vectors for ordering and sequencing;
- Isolation of centromere, telomere, and chromosome-specific clones; and
- Production of large DNA insert libraries of chromosomes

HERAC Report on the DOE Human Genome Initiative

Interagency workshops on resources and informatics initiated

Designation of Lawrence Berkeley Laboratory and Los Alamos National Laboratory as Centers

Initiative management and program plans finalized

Special peer review panels for research proposals

1988 primary R&D awards

Small Business Innovative Research (SBIR) awards

DOE/NIH Memorandum of Understanding

Human Genome Coordinating Committee formed

Human Genome Management Information System initiated

1989 primary R&D awards

SIBR awards

DOE/NIH Five-Year Plan Workshop

Program plan update for 1990

1990 program announcements

DOE/NIH working groups for informatics, mapping, and ethical issues

Contractor-Grantee Workshop

Large Insert Cloning Workshop

Human X Chromosome Workshop

DOE/NIH Five-Year Plan for the Human Genome Project

1990 primary R&D awards

1986 1987 1988 1989 1990

Preface

This nation's Human Genome Project is the first broadly based organized endeavor in the biological sciences. Conceived in 1986, the program was initiated in 1987 as the Human Genome Program of the U.S. Department of Energy (DOE) Office of Health and Environmental Research (OHER). Since that time, it has grown significantly; in 1988 the National Institutes of Health also initiated a human genome program. An ambitious undertaking spanning the disciplines of biology, chemistry, physics, engineering, mathematics, and information science, the national project has a well-defined, long-range endpoint: to decipher the genetic code in the DNA of the entire human genome. Having gathered support in Congress, the Executive Office, and the scientific and commercial sectors, the project of mapping and sequencing the genome has the momentum to make major advances in the knowledge and technologies that are needed to understand the complexities of human cellular processes in a manner never before possible. These advances will impact biological principles as well as the practice of medicine, the growing biotechnology industry, and society.

The project is unusual in that few existing strategies and technologies can be used to achieve its goals. Indeed, the driving force within this endeavor is the development and implementation of innovative, cost-effective methods and technologies for mapping, sequencing, and interpreting the genome. As these developments take place, advances in data analysis and database management will make map and sequence information accessible.

This document is a status report on DOE OHER's Human Genome Program and includes a brief background to this agency's initiative, as well as an explanation of the program's projected focus over the next 15 years. Of special interest are the section on research highlights, the narratives on major genome research efforts conducted at three of DOE's national laboratories, and the abstracts of work in progress. Figures and captions provided by investigators give additional detailed information.

Achievements were reported at the first DOE OHER workshop for grantees and contractors of the Human Genome Program on November 4 and 5, 1989, in Santa Fe, New Mexico. The presentations demonstrated that progress is being made in physical mapping, DNA sequencing, and informatics. DOE plans to convene these workshops on a continuing basis every 18 to 24 months. In the interim, this report and future revisions, together with the bimonthly newsletter and special technical overviews will provide both the interested scientist and layperson with information on developments in this rapidly moving, multidisciplinary project.

Benjamin J. Barnhart

Benjamin J. Barnhart, Manager
Human Genome Program
Office of Health and Environmental Research
Office of Energy Research
U.S. Department of Energy

Acknowledgements

The Office of Health and Environmental Research gratefully acknowledges the contributions made by genome research grantees and contractors in submitting abstracts, photographs, captions, and narratives. Charles Cantor and Sylvia Spengler of Lawrence Berkeley Laboratory provided the material adapted for the "Primer on Molecular Genetics" in Appendix A. "The Alta Summit, December 1984," by Robert Mullan Cook-Deegan was reprinted with the permission of *Genomics*, which is published by Academic Press, Inc. The glossary in Appendix C was adapted from the glossary in the April 1988 U.S. Congress, Office of Technology Assessment publication: *Mapping Our Genes—The Genome Projects: How Big? How fast?* Finally, the Human Genome Management Information System at Oak Ridge National Laboratory (operated by Martin Marietta Energy Systems, Inc., for the U.S. Department of Energy under contract DE-AC05-84OR21400) is recognized for collecting and organizing the information, preparing the manuscript, and implementing the design and production of this publication.

Contents

Human Genome
1989–90 Program Report

Contents

Introduction to the DOE Human Genome Program

The structural characterization of genes and the elucidation of their encoded functions have become a cornerstone of modern health research, biology, and biotechnology. A genome program is an organized effort to characterize all the genetic material—DNA—of an organism. The human genome encodes 50,000 to 100,000 genes on its 24 distinct chromosomes, but only some 2000 genes are now available for study in the form of identified cloned DNAs. To accelerate effective access to all human genes and eventually to generate reference DNA sequences of the chromosomes, the U.S. Department of Energy (DOE) began a Human Genome Initiative in 1986. Intensive studies of the needs and promise of genomics research then ensued. The value of a broad, supportive infrastructure for genomics has now been recognized in the United States and abroad. Genome projects on several important organisms are now planned or are progressing. Two federal agencies are now supporting an expanded national human genome project, and the coordination of human genome research has become international in scope.

Sequencing Technologies

Resonance ionization mass spectrometer. Investigators are shown with the resonance ionization mass spectrometer that is used to measure stable isotopes of a variety of elements that are attached to DNA. Since 50 or more stable isotopes are available for such DNA labeling, a multiplex procedure is available in which either the single radioisotope or the four fluorescent labels are replaced. When performing sequence analysis by using four stable isotopes simultaneously on the four dideoxynucleotide-terminated DNA fragments, the gel lanes needed for electrophoresis can be reduced from the usual four lanes to one lane. Many such elements can be used simultaneously in the same electrophoresis lane since the resonance ionization mass spectrometer will sort out the elements and all their isotopes. Even greater multiplexing would occur when Church probes are used to locate DNA fragments, since all hybridization could be performed simultaneously. Studies are under way to label DNA with a variety of elements that have multiple stable isotopes and to detect the DNA so labeled. These studies are being carried out at the Oak Ridge National Laboratory in collaboration with Atom Sciences, Inc., where this mass spectrometer has been developed to its present state. (Photograph provided by Bruce Jacobson, Oak Ridge National Laboratory, and Heinrich Arlinghaus, Atom Sciences, Inc.)

Introduction

The Office of Health and Environmental Research (OHER) manages the Human Genome Program within DOE, as a focused program of resource and technology development. The general objective is to advance, as economically and efficiently as possible, the U.S. effort in human genome research. Other OHER responsibilities, such as assessing the effects of energy by-products on our population and environment, will benefit from applications of the new biological and computational resources and the innovative technologies developed within the genome program. Moreover, the resources and technologies being developed are of broad and continuing value; many facets of biomedical research, modern agriculture, and the growing biotechnology industries will be advanced.

Knowledge of principles guiding the three-dimensional structure-function relationships of cellular macromolecules is essential for the interpretation and application of the linear DNA sequence that is being elucidated in the Human Genome Program. Special facilities in DOE laboratories are being used (see table on p. 3) to determine the three-dimensional structure of macromolecules. DOE is planning for the greatly increased demand that will be placed on these facilities as a result of the Human Genome Program.

Development of the Human Genome Initiative

In 1984 DOE and the International Commission for Protection Against Environmental Mutagens and Carcinogens cosponsored a workshop in Alta, Utah. A specific charge to the participants was to evaluate the present state of and project future directions for mutation detection and characterization. The growing roles of novel DNA technologies in diagnostics were highlighted. There, in the excitement of the meeting, some core techniques of current genomic analysis were conceived (see "The Alta Summit, December 1984," in Appendix B). These new approaches increasingly included gene cloning and sequencing. Although the isolation of genes from libraries of clones had been an integral component of biomedical research for many years, the one-gene-at-a-time procedures being employed were wasteful of scientists' time and research resources.

The small genomes of several viruses were the targets of the first genome projects in the 1960s. These projects initiated the development of some of the current important techniques in molecular genetics. (A molecular genetics primer is included as Appendix A of this report.) With the advent of molecular cloning techniques in the 1970s, a library of manageable cloned DNAs could be produced for any species. Genome studies on many viruses, the bacterium *Escherichia coli*, two yeast species, and a nematode (a minute worm) were subsequently implemented. With the skills already demonstrated, a human genome program could then be considered. However, it would be a task far more vast than any previously implemented in biological research.

In 1985 DOE began to consider whether its expertise with high-technology projects could facilitate and sustain a human genome program. To assess the desirability and

feasibility of ordering and sequencing clones representing the entire human genome. DOE sponsored in March 1986 an international meeting in Santa Fe, New Mexico. With virtual unanimity the participating experts concluded that this objective was meritorious and obtainable and that it would be an outstanding achievement in modern biology.

The HERAC Recommendation

Further guidance was sought from DOE's Health and Environmental Research Advisory Committee (HERAC), which provided its report on the Human Genome Initiative in April 1987. This report urged DOE and the nation to commit to a large, multidisciplinary, technological undertaking to order and sequence the human genome. This effort would first require significant innovation in the general capability to manipulate DNA. Also required would be major new analytical methods for ordering and sequencing DNA segments, theoretical developments in computer science and mathematical biology, and great expansion in the ability to store and manipulate the information and interface it with other large and diverse genetic databases. The report further recommended that DOE have a leadership role because of its demonstrated expertise in managing complex, long-term multidisciplinary projects, involving both the development of new technologies and coordination of efforts among industries, universities, and its own laboratories.

The role of the Office of Health and Environmental Research (OHER) in its mission to understand the health effects of radiation and other by-products of energy production was noted in the report. This mission requires fundamental knowledge of the effects of damage to the genome, and it has already led to a number of research and technological developments that are integral components of the human genome mapping and sequencing project: DOE computer and data management expertise initiated and supports the GenBank® DNA sequence repository (cosponsored by the National Institutes of Health); the chromosome-sorting facilities essential to the Genome Initiative were developed and are maintained at DOE laboratories; and within the National Laboratory Gene Library Project, libraries of

Major DOE Facilities and Resources Relevant to Molecular Biology Research

Center for X-Ray Optics	**LBL**
GenBank® Data Sequence Repository	**LANL**
High Flux Beam Reactor	**BNL**
Los Alamos Neutron Scattering Center	**LANL**
Molecular Sciences Research Center	**PNL**
National Flow Cytometry Resource	**LANL**
National Laboratory Gene Library Project	**LANL, LLNL**
Protein Structure Data Bank	**BNL**
National Synchrotron Light Source	**BNL**
Scanning Transmission Electron Microscope Resource	**BNL**
Scanning Tunneling Microscopy	**LLNL, ORNL**
Stanford Synchrotron Radiation Laboratory	**Stanford**

Developing Facilities:

Advanced Photon Source	**ANL**
Advanced Light Source	**LBL**

cloned sequences from single human chromosomes are produced for the research community. Thus, the Human Genome Initiative was a natural outgrowth of ongoing DOE-supported research.

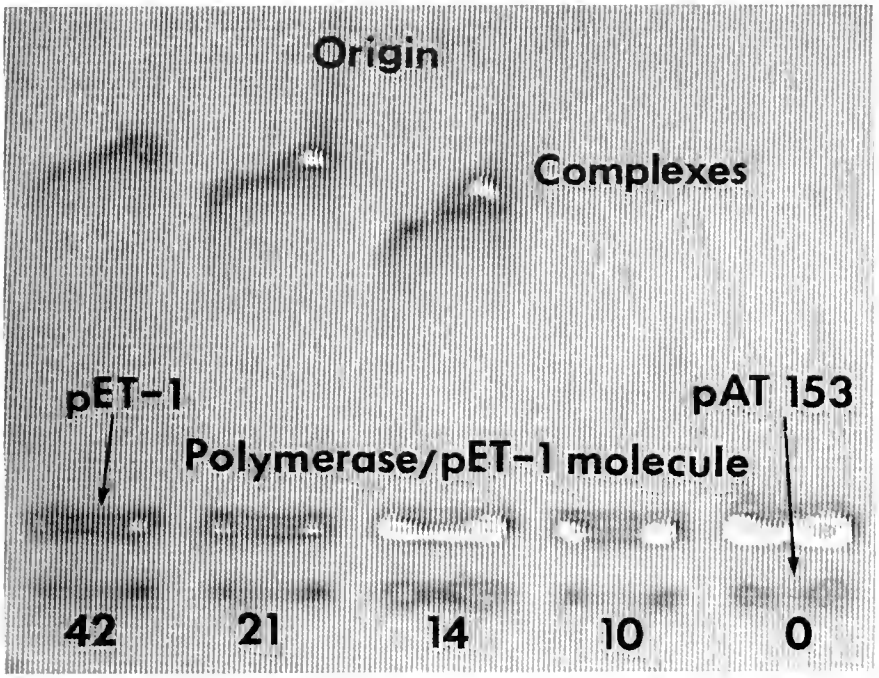
OHER responded to the Santa Fe meeting and HERAC reports by implementing three major objectives that are being pursued concurrently:

1. The generation of refined physical maps of the chromosomes, including the ordering of representative libraries of DNA clones.
2. The development of requisite supportive strategies, chromosomal resources, and instrumentation, which includes development and testing of advanced sequencing technologies.
3. The expansion of communication networks and computational and database capacities and the development of advanced algorithms for managing and interpreting the clone ordering and sequence data.

A small number of genes or other selected regions of interest will be sequenced during the generation of physical maps; however, the transition from mapping to an intensive genome-sequencing effort awaits development of more accurate, rapid, and economical technologies that are needed to commence large-scale sequencing. Once suitable new technologies are implemented, contiguous segments of DNA will be decoded into a reference sequence of the human genome. These segments will be derived from ordered clones and DNA fragments that are identified or mapped to particular locations on chromosomes by sequence-tagged sites (STSs) or other methodologies. Emerging methodologies will be tested and validated in pilot sequencing projects prior to incorporating any single protocol as the primary method.

As implementation of the OHER program began with a small number of pilot projects, other government agencies, scientific societies, and commercial organizations initiated their own studies of associated policy and strategy issues and presented their recommendations. The most prominent reports are those of the National Research Council (NRC) and the Congressional Office of Technology Assessment (OTA). While broadly in accordance with the earlier HERAC recommendations, the NRC and OTA reports further recommended that several nonhuman species also be included in the national effort and that the physical mapping of chromosomes be complemented by genetic mapping.

The OHER human genome program remains focused on physical map construction and development of advanced sequencing methodologies and technologies. Because many of the resources and technologies being developed have broad applicability, they contribute substantially to OHER programmatic objectives in the fields of radiation biology, chemical toxicology, molecular epidemiology, and the ecological and environmental biosciences and aid the developing genome programs of other agencies as well.



Mapping Instrumentation

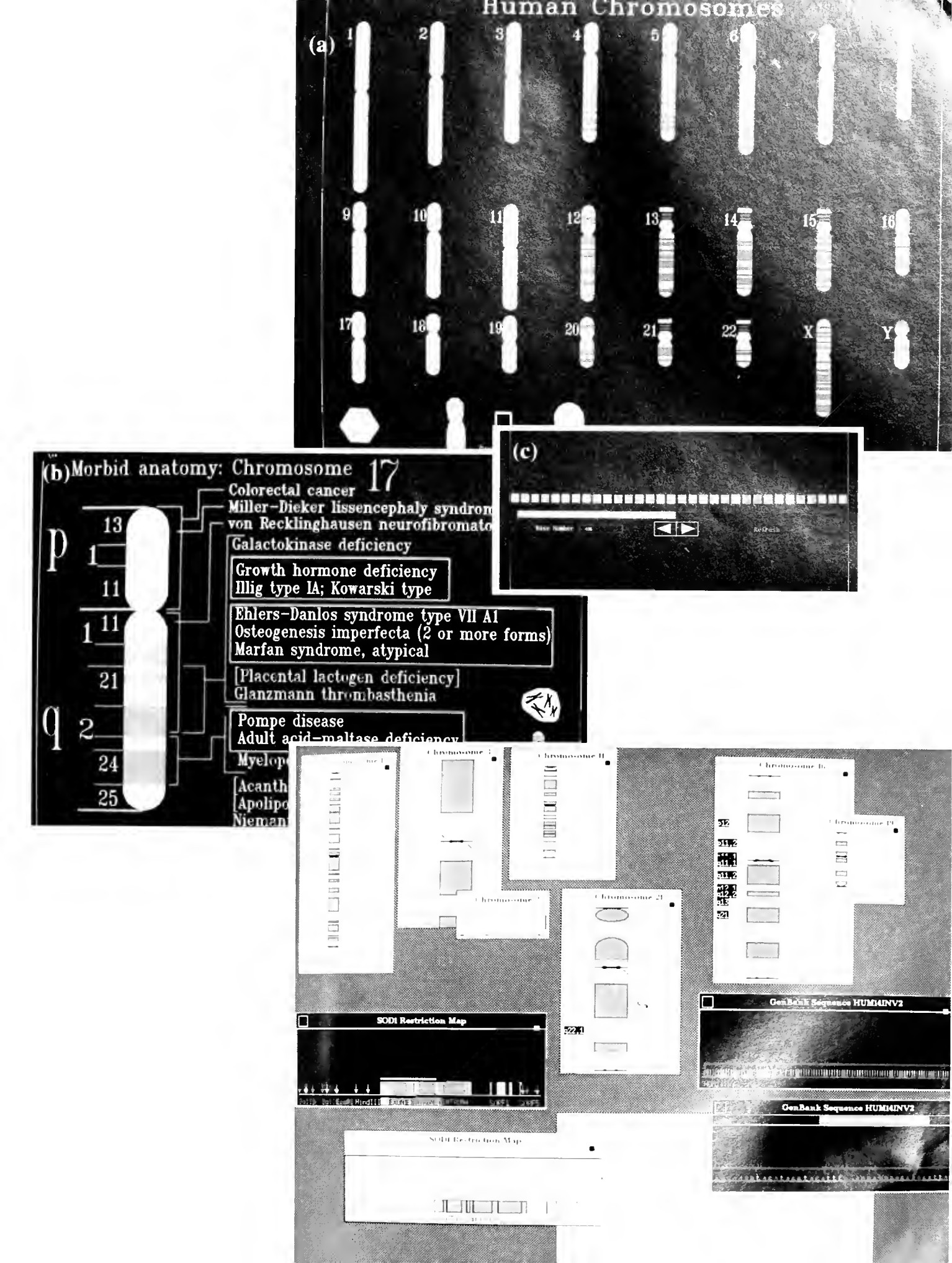
DNA-Protein-Binding Assay — Use of gel retardation for recognition of promoter sequences that are necessary for the polymerase enzyme to synthesize DNA for sequencing studies. Increasing quantities of T7 RNA polymerase were incubated with pET-1 DNA (contains a strong T7 promoter) and pAT153 DNA (no promoter) for 10 min at room temperature; samples were then electrophoresed in a 1% agarose gel for 2 hr at 2.5 V/cm. As the ratio of polymerase molecules to DNA increased, the quantity of pET-1 DNA band decreased, and a new band with lower mobility appeared. Without requiring quantitative loading of the gel, the ratio of fluorescence in the pET-1 band to that in the control pAT153 band permits quantitation of the fraction of the pET-1 molecules bound to polymerase. This is a useful technique for finding specific promoter sequences; once found, these promoter sequences can be attached to DNA fragments of choice (e.g., fragments the researcher is interested in sequencing). (Photograph provided by Betsy Sutherland, Brookhaven National Laboratory.)

Informatics

Data access through interactive workstations. One of the great challenges of the human genome project is how to integrate and provide access to the growing mass of genomic data. One solution is development of sophisticated workstations that would provide a uniform user interface with all map and sequence databases. With the prototype developed at Lawrence Berkeley Laboratory's Human Genome Center, the user examines data at increasing resolution by "enlarging" selected regions of successive displays. The three illustrations shown here display **(a)** the full complement of chromosomes, **(b)** a single chromosome with locations of human disease genes, and **(c)** the nucleotide sequence for a selected region within that chromosome. Within this region, the order of the nucleotide bases is displayed by the following colors: A (chartreuse), C (orange), G (light blue), and T (pink). The icons at the bottom or side of figures a–c indicate access to other levels of information about each chromosome including staining, gene mapping, morbidity (disease), and sequence. In figures a and b, the dark blue bands indicate the characteristic Giemsa staining patterns; the chromosome centromeres are pink; and the yellow areas on the chromosomes represent the heterochromatic regions (C–banding pattern). Less characterized areas are light blue. (Photographs a–c provided by the Human Genome Center, Lawrence Berkeley Laboratory.)

Informatics

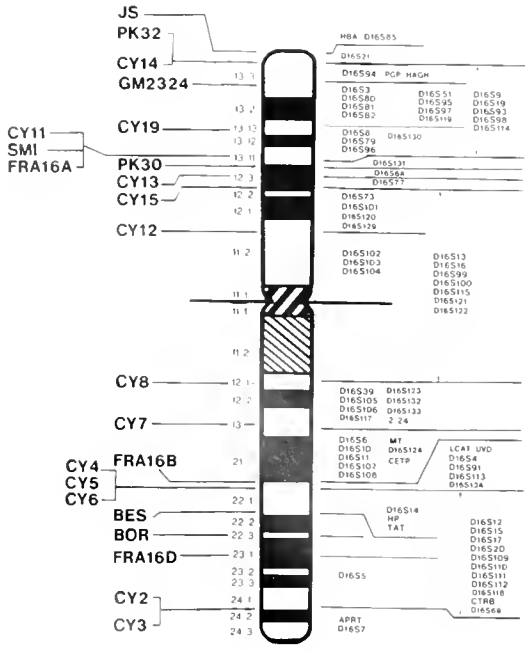
GnomeView Workstation. The mosaic shown in the bottom photograph illustrates the versatility of the X-window system that is part of the GnomeView Interface currently in use at the U.S. Department of Energy's Pacific Northwest Laboratory. Shown on the same screen are simultaneous views of chromosomes at various magnifications (upper screen), restriction maps (windows on lower left), two magnification levels of a sequence from GenBank® (windows on lower right), and a GenBank® file information header (text window, lower screen). The X-window system, coupled with the network model database system of the GnomeView Interface, allows easy access and simultaneous viewing of information from all levels of the human genome hierarchy. (Photograph provided by Richard Douthart, Pacific Northwest Laboratory.)


(a)
Human Chromosomes
1
2
3
4
5
6
7
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
(b) Morbid anatomy: Chromosome 17
Colorectal cancer
Miller-Dieker lissencephaly syndrom
von Recklinghausen neurofibromato
Galactokinase deficiency
Growth hormone deficiency
Illig type IA; Kowarski type
Ehlers-Danlos syndrome type VII A1
Osteogenesis imperfecta (2 or more forms)
Marfan syndrome, atypical
[Placental lactogen deficiency]
Glanzmann thrombasthenia
Pompe disease
Adult acid-maltase deficiency
p
q
13
1
11
11
21
2
24
25
(c)
Refresh
SOD1 Restriction Map
GenBank Sequence HUMI4INV2
GenBank Sequence HUMI4INV2

OHER Mission

The Office of Health and Environmental Research (OHER) has research and development responsibilities that are mandated by 1946 and 1954 legislative acts (see "Enabling Legislation" on p. 10), some of which have been carried forward from DOE's predecessor agencies, the Atomic Energy Commission (AEC) and the Energy Research and Development Administration (ERDA). The first national support for genetics research was provided by AEC; further responsibilities were authorized in 1974 and 1977. Although the initial focus was on radiation effects, the objectives were later broadened to include the health consequences of all energy technologies and their by-products. Long-range goals are to address applications of the resources and technologies developed in the genome program to the Department's interests in genetic damage from exposures to ionizing radiation and chemicals. An extensive program of OHER-sponsored research on genome structure, maintenance, damage, and repair continues at the national laboratories and universities.

A major concern today is human exposure to background environmental factors and how the body responds to such factors. In the environment there are unavoidable genome-damaging agents from which we are at risk. Among them are natural radiation sources, which include components of sunlight, cosmic rays from space, and the radon released from the Earth. There are both inorganic and organic chemicals that can cause DNA damage. Some of these chemicals are natural to the environment, while others are generated by human commerce and energy-related processes.



Physical Mapping

Diagram of human chromosome 16 showing the G-banding (Giemsa-staining) pattern. On the left of the figure are the points at which chromosome breaks have been defined. A correlation has been found between the occurrence of these breaks and other chromosome anomalies, such as translocations, deletions, or fragile sites (the latter designated by prefix FRA). Mouse/human somatic hybrids (designated by prefix CY) have been constructed by transferring a portion of chromosome 16 to a mouse cell line. On the right of the figure are names of cloned DNA fragments (probes) from human chromosome 16. The fragments have been mapped to the defined regions of this chromosome by Southern blot analysis of DNA from the somatic cell hybrids and by in situ hybridization. The DNA probes are either anonymous cloned fragments of DNA or cloned genes. (Figure provided by Grant Sutherland, North Adelaid Children's Hospital, Australia.)

Normal biological activities also contribute to the risk of genetic damage. A body's own cells produce some potentially damaging molecules in the course of normal metabolic processes; some of these molecules are produced in considerable abundance during defensive actions against microbes and during detoxification of harmful environmental substances. The genome replication system itself sometimes errs during cell proliferation. Even DNA is not completely stable chemically; its normal methyl-cytosine constituent has a low but measurable rate of spontaneous mutagenic change.

Life has thus evolved under a continuous low-level infliction of genomic damage and mutation. Under this pressure, systems that reverse or ameliorate many types of DNA damage have evolved, so that a wide range of repair mechanisms exists within cells of all species. Several of the human genes contributing to DNA repair processes are being characterized now, and others await detection and molecular cloning. Repair gene deficiencies are manifested as cellular sensitivity to low-level DNA damage and in diseases such as cancer. Humans exhibit genetic diversity in capacity for DNA repair in response to ubiquitous DNA-damaging agents.

In recognition of this diversity, a major goal of the current OHER health effects and general life sciences program areas has been formulated: the development of capacities to diagnose individual susceptibility to genome damage imposed by energy-related factors. Some major components of this OHER research are:

- molecular cloning and characterization of DNA repair genes;
- improvement of methodologies and development of new resources for use in quantitating and characterizing mutations (molecular epidemiology); and, most recently,
- focused resource and technology development needed to map and sequence the human genome—the Human Genome Program.

Enabling Legislation

The Atomic Energy Act of 1946 (P.L. 79-585) provided the initial charter for a comprehensive program of research and development related to the utilization of fissionable and radioactive materials for medical, biological, and health purposes.

The Atomic Energy Act of 1954 (P.L. 83-703) further authorized the AEC "to conduct research on the biologic effects of ionizing radiation."

The Energy Reorganization Act of 1974 (P.L. 93-438) provided that responsibilities of the Energy Research and Development Administration (ERDA) shall include "engaging in and supporting environmental, biomedical, physical and safety research related to the development of energy resources and utilization technologies."

The Federal Non-nuclear Energy Research and Development Act of 1974 (P.L. 93-577) authorized ERDA to conduct a comprehensive non-nuclear energy research, development, and demonstration program to include the environmental and social consequences of the various technologies.

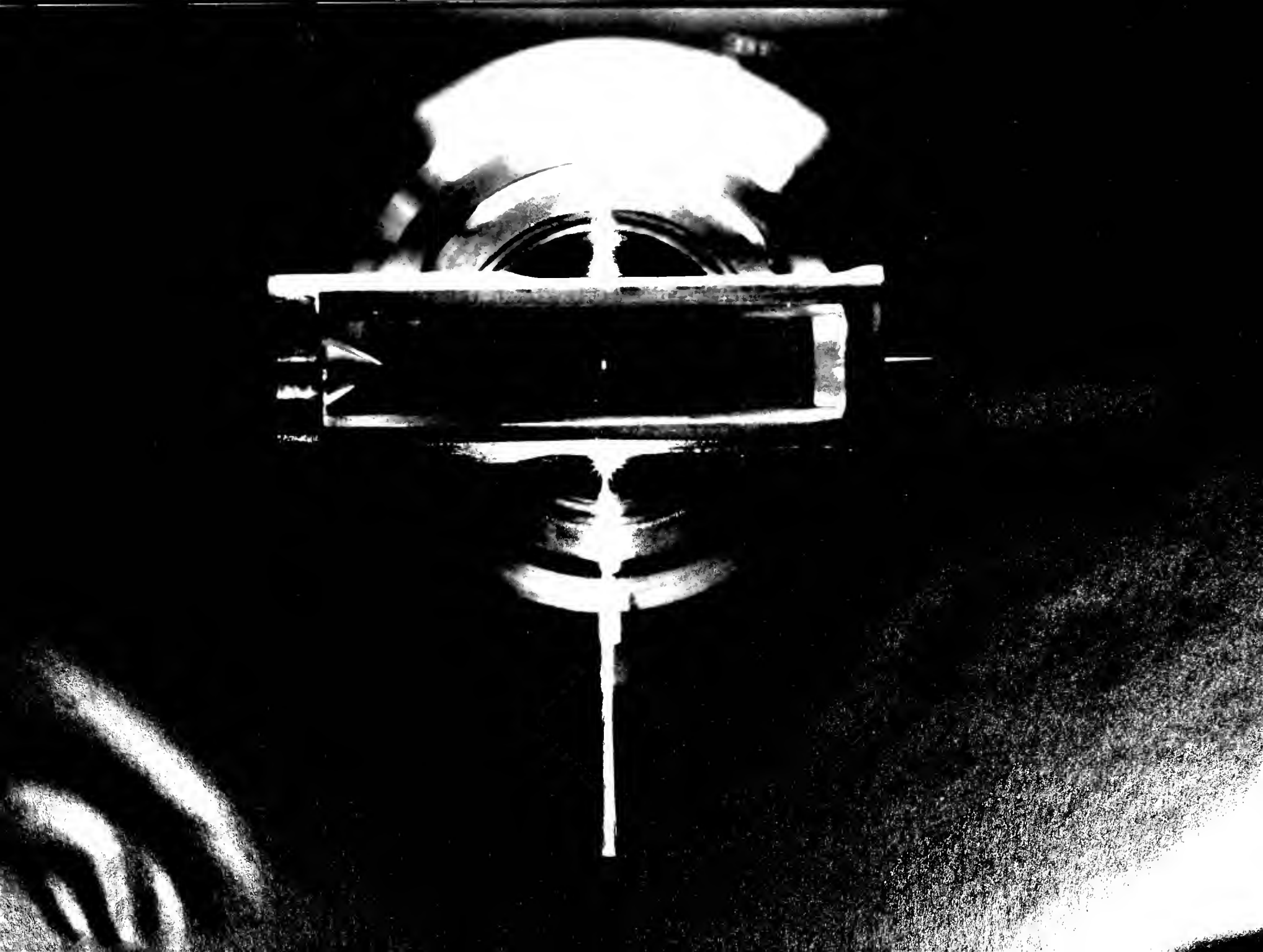
The DOE Organization Act of 1977 (P.L. 95-91) mandated the Department "to assure incorporation of national environmental protection goals in the formulation and implementation of energy programs; and to advance the goal of restoring, protecting, and enhancing environmental quality, and assuring public health and safety," and to conduct "a comprehensive program of research and development on the environmental effects of energy technology and program."

Physical Mapping

Researchers comparing photographs of gels in which restriction fragments have been separated. The technology being developed includes a reliable method for producing a partial digest of DNA in agarose. (Photograph provided by Michael McClelland, California Institute of Biological Research.)

Sequencing Technologies

Application of flow cytometry to DNA sequencing. The small yellow spot in the center of this multiple-exposure photograph shows the fluorescence from approximately 1000 molecules of the laser dye—rhodamine 6G. The apparatus shown is a modified flow cytometer with the green argon laser beam traversing from left to right; the flow cuvette is vertical in the center. The fluorescence collection lens can be seen in the background. An apparatus similar to the one shown in the photograph is being developed to sequence DNA by detection of single, fluorescent molecules. (Photograph provided by James Jett, Los Alamos National Laboratory.)



Management of the Human Genome Program

Program Management Structure

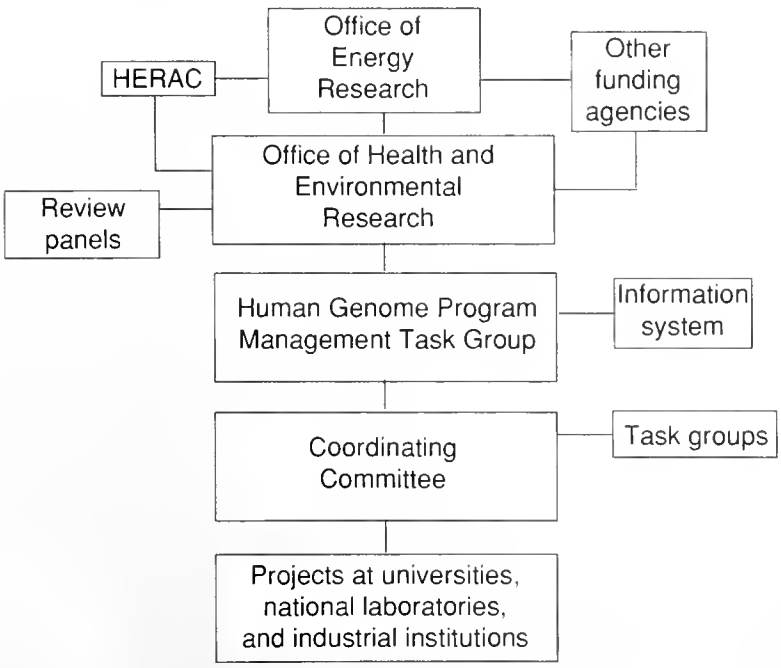
The highly multidisciplinary, focused, and long-term character of the Human Genome Program is novel to biological research. An infrastructure connecting biomedical research, technology development, computer sciences, data and physical repositories, and supporting agencies has thus become essential. The Health and Environmental Research Advisory Committee (HERAC) provides policy, strategy, and scientific guidance to the program.

Program Management Task Group

Within DOE, the management structure recommended by HERAC is that the Human Genome Program Manager and Management Task Group work within the Office of Health and Environmental Research to coordinate:

- Independent scientific boards that provide peer review of research proposals; both prospective and retrospective evaluations are utilized.
- Administration of awards, collaboration with all concerned agencies and organizations, organization of periodic workshops, and responses to the needs of the developing program.
- The support services provided by the Human Genome Management Information System (HGMIS) at Oak Ridge National Laboratory. As a DOE management tool to facilitate communications among management and research personnel and to update the public on genome research, HGMIS publishes newsletters and technical and other program reports and maintains an electronic bulletin board that carries current information and announcements. The on-line bulletin board and publications are available to all persons interested in the genome project.

DOE Human Genome Program Management and Coordination



Human Genome Coordinating Committee

Another component of the DOE management structure is the Human Genome Coordinating Committee (HGCC), which was chartered by HERAC. The committee, originally named the Human Genome Steering Committee, was formed in October 1988. The HGCC membership comprises and represents DOE genome program research participants. Members of the Human Genome Program Management Task Group (ex-officio members of the HGCC) and observers from other government and private agencies participate in the regularly scheduled meetings of the HGCC, whose responsibilities include:

- assisting OHER with the overall coordination of DOE-funded genome research,
- facilitating the development and dissemination of novel genome technologies,
- ensuring proper management of data and samples,
- interacting with other national and international efforts,
- communicating the program to the press and public, and
- establishing task groups to analyze specific issues such as ethics, informatics, resource sharing, cost of resource distribution, and use of chromosome-flow-sorting facilities.

Human Genome Coordinating Committee

Chairman: Charles R. Cantor, Director, Human Genome Center, *Lawrence Berkeley Laboratory*

Anthony V. Carrano, Director, Human Genome Project, *Lawrence Livermore National Laboratory*

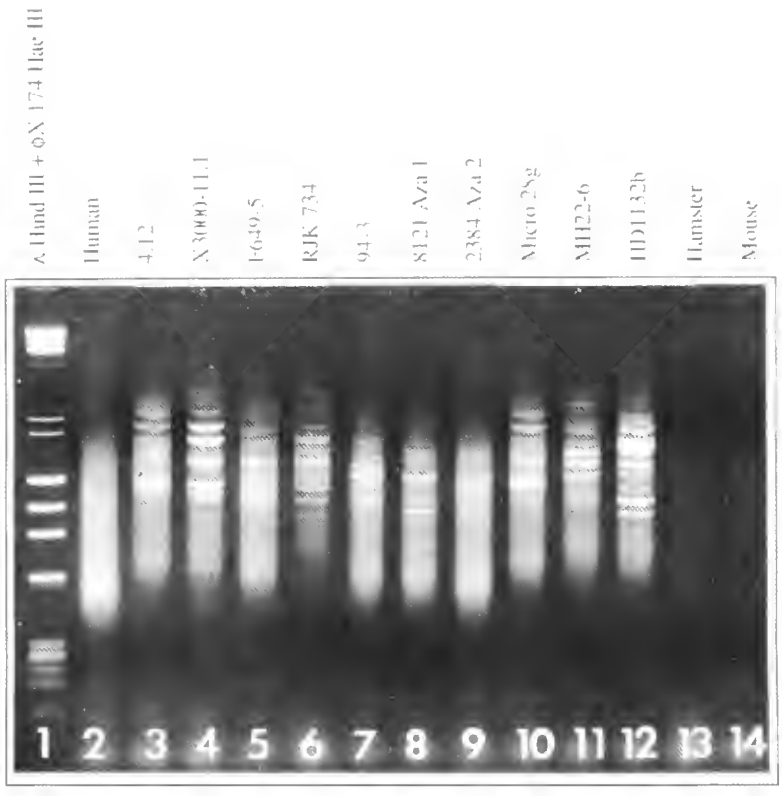
C. Thomas Caskey, Director, Institute for Molecular Genetics, *Baylor College of Medicine*

Leroy E. Hood, Director, Center for Integrated Protein and Nucleic Acid Chemistry and Biological Computation, and Director, Cancer Center, *California Institute of Technology*

Robert K. Moyzis, Director, Center for Human Genome Studies, *Los Alamos National Laboratory*

HGCC Executive Officer: Sylvia J. Spengler, *Lawrence Berkeley Laboratory*

Management of the Human Genome Program



TC-65

Physical Mapping

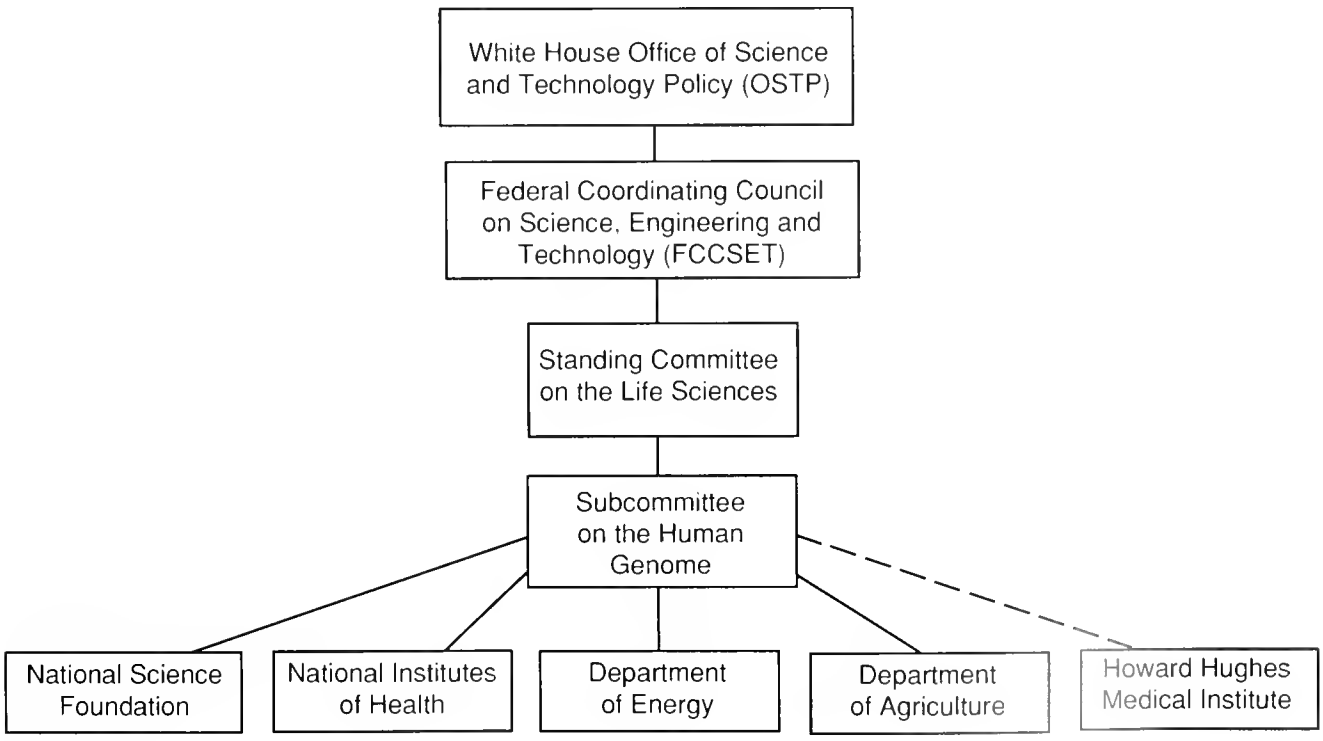
Human DNA fragments obtained from rodent somatic cell hybrid background separated on agarose gels. A primer designed to recognize human *Alu* sequences is used for rapid amplification of regions of human DNA in rodent/human somatic cell hybrids. Rodent/human hybrid cells are constructed and used in human genome studies because they contain manageable amounts of human DNA in which genome regions of interest can be manipulated and characterized. The TC-65 oligonucleotide primer was designed to recognize the human, but not the rodent, *Alu* sequences and provides specific amplification of human DNA between regions of these ubiquitous *Alu* sequences when polymerase chain reaction (PCR) methods are used. (*Alu* sequences are about 300 bp long and repeated thousands of times in the human genome.) The specificity of the TC-65 primer is demonstrated in the figure: DNA fragments of total human genome (lane 2) and fragments of different rodent/human hybrid cell lines have been amplified and separated by gel electrophoresis (lanes 3–12). Note the abundance of bands (white) of DNA fragments in lanes 2–12 and the lack of fragments in lanes 13 and 14, where pure rodent genome samples were electrophoresed. No human *Alu* repeat sequences are found in the rodent genomes, and the rodent *Alu* equivalent sequences are not amplified; this TC-65 primer/PCR method is thereby validated. Lane 1 contains standard DNA fragments of known size for determining sizes of DNA fragments in the other lanes.

This method is useful for rapid comparison of hybrid cell lines' DNA content and overlap and can also be used in preparation of nucleic acid probes from cloned human DNAs—especially for clones in yeast artificial chromosome (YAC) vectors. [Photograph was first published in *Proc. Natl. Acad. Sci. USA* **86**, 6686–6690 (1989). Photograph provided by David L. Nelson and C. Thomas Caskey, Baylor College of Medicine.]

Interagency Coordination

The U.S. agencies engaged in genome research meet formally under the auspices of the White House Office of Scientific and Technology Policy. The Department of Agriculture is initiating a genomics program. The National Science Foundation has computational and informatics programs supportive of genomics in addition to individual awards in genetics and molecular biology. The Howard Hughes Medical Institute, a private foundation, contributes substantially to the genome effort through its support of biomedical research and related infrastructure. The National Institutes of Health (NIH) started its own genome program in 1988. The NIH program complements the DOE program by supporting predoctoral and postdoctoral training in molecular genetics, studying model organisms, emphasizing genetic diseases, and preparing human genetic maps requiring family studies.

Interagency Coordination of Genome Research



Interagency Coordination

Management of the Human Genome Program

Joint DOE/NIH Activities

A 1988 Memorandum of Understanding specifies procedures for coordinating DOE and NIH efforts and establishes a joint advisory committee and an interagency working group. NIH observers attend the quarterly meetings of the DOE Human Genome Coordinating Committee, and DOE observers attend meetings of the Program Advisory Committee (PAC) to the NIH National Center for Human Genome Research. In August and October, DOE and NIH representatives met informally as a joint planning group to begin formulation of a coordinated multiyear research plan, which was presented in December to the HERAC and PAC subcommittees, who then completed the plan. This national plan was presented to the U.S. Congress in early 1990.

Several important workshops have been cosponsored by DOE and NIH; they include:

- Workshop on Repositories, Data Management, and Quality Assurance for the National Gene Library and Genome Ordering Projects (August 1987),
- Workshop on Data Management for Physical Mapping (cosponsored with the Howard Hughes Medical Institute) (May 1988),

Informatics

Human Genome Management Information System (HGMIS) staff. HGMIS staff are located at the Oak Ridge National Laboratory in the Biomedical and Environmental Information Analysis Section of the Health and Safety Research Division. Members of the Graphics Division and the Publications Division assist with manuscript design and preparation. HGMIS welcomes contributions and suggestions from genome researchers. (Photographs provided by HGMIS, Oak Ridge National Laboratory.)

- Workshop on Nomenclature for Physical Mapping of Complex Genomes (cosponsored with Howard Hughes Medical Institute) (April 1989),
- Large Insert Cloning Workshop (December 1989), and
- Workshops on Chromosomes 16 and X (June and December 1989).

The Joint Informatics Task Force, comprised of experts appointed by the NIH PAC and the DOE HGCC, has constructed a document that makes recommendations for the present and future computing needs of researchers involved in the Human Genome Project. Another joint DOE/NIH working group has been formed to address ethical, social, and legal issues associated with the Human Genome Project. A third joint working group, on chromosome mapping, is being formed.

International Human Genome Organisation

The international Human Genome Organisation (HUGO) has been formed to assist with coordination of national efforts, facilitate exchange of research resources, encourage public debate, and provide information and advice on the implications of human genome research. Conceived in April 1988, HUGO is incorporated in Switzerland and in the United States. New members from among participants in genome research are elected in a manner similar to that of the European Molecular Biology Organisation and in some ways parallel to the U.S. National Academy of Sciences. Its 42 founding members represented 17 countries and included 3 members of the DOE HGCC. Within its first year, 219 members were elected, including 12 participants in DOE-funded genome projects. The election of 20 new members in December 1989 increased the membership to 239. HUGO's officers for 1990 are Sir Walter Bodmer (United Kingdom), President; Charles R. Cantor (United States), Vice President, North America; Kenichi Matsubara (Japan), Vice President, Asia; and Andrei D. Mirzabekov (Soviet Union), Vice President, Eastern Europe.

Resource Allocation

Management of the Human Genome Program

The reports of HERAC and the National Research Council on the Human Genome both recommended that national funding for the Human Genome Initiative increase to reach a sustaining yearly level of $200 million. The expenditures within the DOE program have been $5.5 million in FY 1987, $10.7 million in FY 1988, $17.5 million in FY 1989, and $25.9 million in FY 1990. The presidential budget for the DOE Human Genome Program in FY 1991 is $46.0 million, as shown in the figure. Major administrative categories are:

Resource and Technology Development

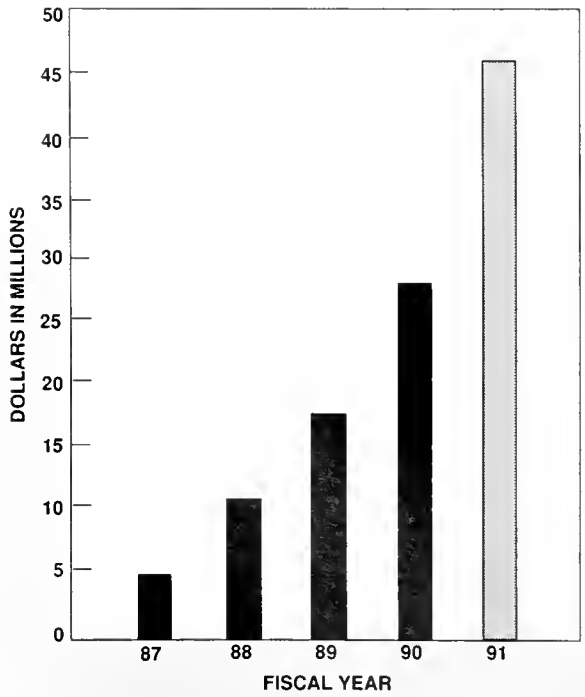
- The National Laboratory Gene Library Project
- Instrumentation and biological support for physical mapping
- Physical mapping through clone ordering and macro-restriction analyses
- Sequencing technology, including automation and robotics
- Data management, analysis, and networking, including GenBank®

Training

- Predoctoral and postdoctoral fellowships at national laboratories

Supporting Activities

- Human Genome Coordinating Committee
- Task groups
- Human Genome Management Information System
- Workshops
- Ethical and societal issues
- Support for national and international meetings
- Improvements to national laboratory resources and facilities
- Technology transfer activities
- Publications



Expenditures and FY 1991 presidential budget for the DOE Human Genome Program

Continuing Implementation

The 1987 HERAC report on the Human Genome Initiative provided the broad guidelines for OHER's Human Genome Program. Refined management and program plans were prepared in 1988. With the experience and progress now achieved and with participation of the Human Genome Coordinating Committee, program guidelines for DOE Human Genome Program implementation have been updated:

Short-Term Focus (1–5 Years):

- Improve by an order of magnitude the efficiency and cost-effectiveness of mapping and sequencing technologies.
- Rapidly develop a database system for current single chromosome projects.
- Complete orderings of monochromosomal clone libraries already initiated.
- Initiate physical mapping of additional chromosomes.
- Improve and implement methods for information and materials dissemination.
- Develop a long-term human genome database system.
- Continue small-scale DNA sequencing as an adjunct to physical mapping, and as a test bed for improved sequencing concepts and technologies.
- Encourage increased private involvement in all areas of genomics.

Mid-Term Focus (5–10 Years):

- Accelerate DNA sequencing as more efficient systems are validated.
- Continue development of algorithms for interpreting sequence information.
- Utilize accumulating genome knowledge to improve assessments of individual susceptibility to genetic damage, from both unavoidable environmental agents and, especially, energy by-products.
- Utilize accumulating genome knowledge to identify the more biologically significant damage sites in chromatin.
- Elucidate the structure, function, and interaction of the body's macromolecules by complementing DNA sequence information with the national laboratories' unique technologies for structural biology studies.

Long-Term Focus (10–15 Years):

- Complete large-scale DNA sequencing and apply interpretative algorithms.
- Emphasize applications of genome knowledge to prospective and retrospective analysis of individual exposures to low levels of energy-related agents.

Physical Mapping

Localization of unique cosmid clones delineates physical landmarks on chromosomes. (a) A powerful approach for constructing physical maps is to use fragments of human DNA cloned in cosmid vectors in in situ hybridization experiments with the chromosomes being mapped. These fragments can be localized to specific regions (accurate to within 1% of the chromosome length) on human metaphase chromosomes by using computer-controlled confocal laser microscopy to detect fluorescence hybridization between the fragments and complementary regions on chromosomal DNA. The researcher in the background is operating the microscope to produce an image on the distant monitor. The researcher in the foreground is performing computer analysis on the digitized hybridization data. **(b)** Shown in this photo is the in situ hybridization of an anonymous fluorescently labeled cosmid to the long arm of human chromosome 11. The chromosomes are stained with propidium iodide (red), and cosmid hybridization is indicated by yellow fluorescence from fluorescein. The red dye on the chromosome is uniform, except at the location of in situ hybridization as indicated by line *f* on the graphic. [Photograph b first published by the American Association for the Advancement of Science in P. Lichter et al., "High-Resolution Mapping of Human Chromosome 11 by in Situ Hybridization with Cosmid Clones," *Science* **247**, 64–69 (Jan. 5, 1990). Photographs a and b provided by Glen Evans, The Salk Institute, and Peter Lichter, Yale University Medical School.]

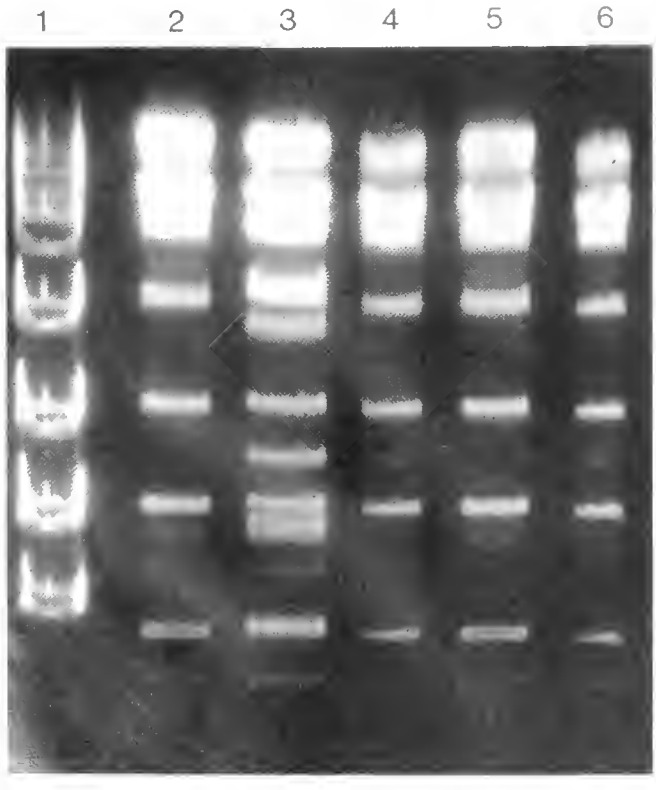

(a)


(b)
p
f

1989–90 Research Highlights

The first research and development projects supported by the Human Genome Initiative were pilot projects in the national laboratories and in academia. Subsequent projects have been initiated after evaluation by special peer review panels in 1988 and 1989. Abstracts of all current projects are included in this report and are supplemented by research narratives from national laboratories and by special reports in the Appendices. There have been numerous incremental contributions to the resource and technology development, in addition to significant progress toward the major goals. Some highlights of the total DOE program include:

The construction of libraries made up of DNA clones with large-capacity phage/cosmids containing human DNA is progressing within the National Laboratory Gene Library Project. These libraries will represent the 24 distinct chromosomes (one chromosome representing each of the 22 autosome pairs plus the X and Y chromosomes) and, even now, are an extremely valuable resource for physical mapping projects. Libraries representing chromosomes 4, 5, 8, 11, 17, 21, and 22 are being offered for evaluation and cooperative use in 1990.

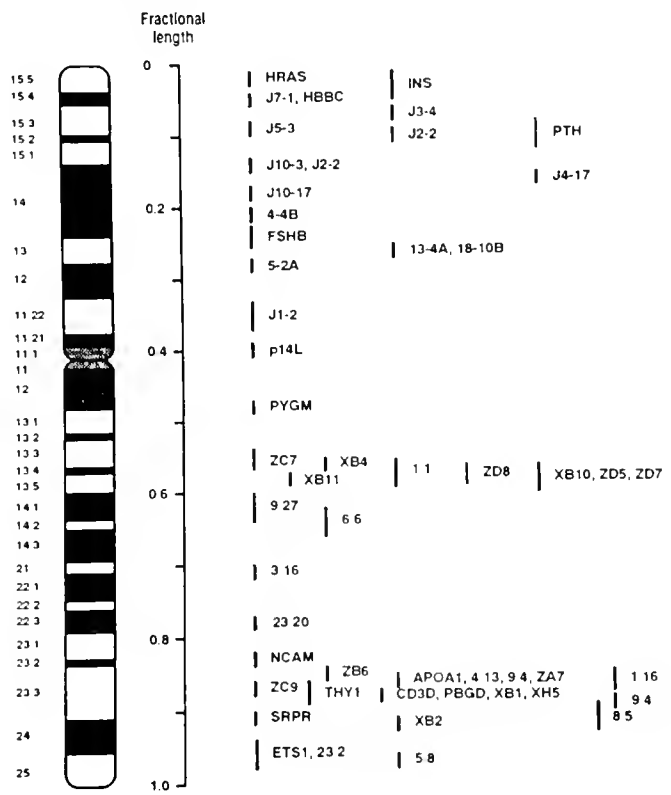


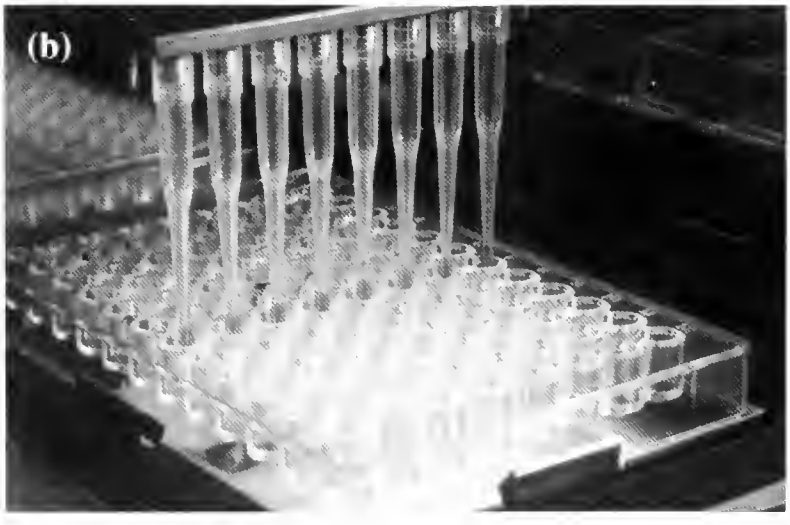
Physical Mapping

Cross-protection against *Not* I sites in *E. coli*. Physical map construction is more efficient when large DNA fragments are utilized. The restriction endonuclease *Not* I cleaves the sequence 5′-GCGGCCGC-3′; however, if DNA is first methylated at ^{m}CGCG with M•*FnuD* II or M•*Bep* I, a subset of the *Not* I sites cannot be cleaved; the specificity of *Not* I is effectively doubled. Shown in the figure is a pulsed-field gel of *E. coli* RRI genomic DNA treated as follows. **Lanes 2 and 5**: Methylated at ^{m}CGCG by M•*FnuD* II, then cut with *Not* I. **Lane 3**: Unmethylated, cut with *Not* I. **Lanes 4 and 6**: Methylated at ^{m}CGCG by M•*Bep* I, then cut with *Not* I. **Lane 1**: Bacteriophage lambda concatemer ladder, molecular weight marker, 48,502 bp per step. (Photograph provided by Michael McClelland, California Institute of Biological Research.)

Physical Mapping

Physical map of contigs on chromosome 11. Researchers at The Salk Institute and Yale University Medical School have generated a series of overlapping sets of cosmids, or contigs, that vary from 2 to 27 clones. The position and relative order of many of the contigs have been determined by using fluorescence in situ hybridization on metaphase chromosome spreads. In particular, the relative order of contigs containing known cloned genes, anonymous DNA markers, or *Hpa*-II-tiny-fragment (HTF) islands (possibly indicating the location of as yet undescribed genes) are indicated for comparison to the ideogram of chromosome 11. The position of hybridization is determined by the fractional length from the 11p telomere (FLpter) rather than using cytogenetic banding. Known genes that have been mapped on the long arm include those encoding the neural cell adhesion molecule (NCAM), the Thy-1 antigen, the ApoA1 cluster, the CD3 cluster, porphyrinogen deaminase (PBG), the ETS1 oncogene, and the signal recognition particle receptor (SRPR), as well as others. [Figure first published by the American Association for the Advancement of Science in P. Lichter et al., "High-Resolution Mapping of Human Chromosome 11 by in Situ Hybridization with Cosmid Clones," *Science* **247**, 64–69 (Jan. 5, 1990). Figure provided by Glen Evans, The Salk Institute, and Peter Lichter, Yale University Medical School.]





Physical Mapping

Automated robotic system used to prepare DNA cosmid clones. (a) The investigator is using a robot to prepare sample solutions that contain cloned DNA fragments for loading onto gels for electrophoresis. Electrophoresis is used to determine the size of cloned DNA fragments prior to employing the fragments in hybridization techniques. **(b)** Close-up view of the eight-channel pipette arm that prepares and loads samples onto gels. (Photographs provided by Glen Evans, The Salk Institute.)

The ordering of DNA clones has been initiated for chromosomes 5, 11, 16, 17, 19, 21, 22, and X. Efficacy and speed have been demonstrated in these projects for three distinct ordering strategies.

The 1988 identification of the basic repeat sequence of the human telomere at LANL has been followed by the cloning of telomeric regions of several chromosomes. Thus the "end points" of the physical mapping tasks are becoming well defined and are providing orientation for mapping activities. The telomeric sequence has been found to be conserved across vertebrate species.

Novel computer software now makes possible the direct entry of raw experimental results into a database, subsequent data analysis, and future transmission of results to other laboratories and data repositories. These systems are simplifying the requirements for recording and processing map and sequence data.

Broader problems in the area of human genome informatics have been addressed in a series of workshops cosponsored by concerned federal agencies. To pursue these issues further, the Joint Informatics Task Force (JITF) has now been formed. Recently, guidelines have been published to eliminate ambiguities in clone names and thus provide for unique naming or identification.

Very fast computer boards for sequence search and comparison tasks have been demonstrated and are being commercialized.

Improvements in protocols for construction of yeast artificial chromosomes (YACs) have culminated with the production of YAC libraries whose human DNA inserts have an average size of 410,000 bp.

For chromosome mapping through pulsed-field gel electrophoresis, the number of useful cleavage sites has been increased by protocols for modifying DNAs in vitro.

The reliability of a core DNA-sequencing step of the Sanger strategy has been substantially increased, through genetic engineering of DNA polymerase (of bacteriophage T7) and modification of polymerase reaction conditions.

A scheme for rational combination of random and directed-sequencing runs on cosmids provides for much more economical use of expensive primers.

The processing of DNA fragment autoradiographs into assembled sequence data is now being accelerated by automatic film readers coupled with computerized analysis.

A novel scheme for sequencing by hybridization (SBH) crucially depends on a capacity to distinguish short segments of perfectly base-paired DNA from segments with even a single base-pair mismatch. Both the effective theory and practice for such discrimination have now been demonstrated.

Research Highlights

Instrumentation and methodologies, being developed to use multiple stable isotopes as DNA labels for mapping and sequencing tasks, will increase the speed of sequencing.

Chemiluminescent techniques for displaying DNA fragments are now achieving sensitivities of radioisotopic labels, with promise for reduction in safety hazards and hazardous waste disposal costs.

The first nondestructive images of naked DNA have been achieved through scanning tunneling microscopy and provide promise for single-molecule DNA sequencing.

A multiplex walking strategy has been applied to an 1100-member library representing chromosome segment 11q. Automated restriction mapping is now being utilized to confirm/reject the presence of overlaps between cosmid pairs with homologies. Some 300 contigs have thus far been constructed.

An automated fluorescence-based method for clone fingerprinting has been developed, validated, and coupled to software used for contig assembly, data storage, and graphical display of map information. These procedures are being successfully applied toward the development of a cosmid and YAC contig map of human chromosome 19.

A new method of genome mapping using human-specific repetitive sequences and the polymerase chain reaction (*Alu* PCR) has been developed. This method is used for the isolation of regionally localized DNA fragments and for the rapid and efficient characterization of cloned DNAs, particularly those in YAC vectors.

Resource Development

Computerized robotics used to speed repetitive tasks of mapping and sequencing DNA. Application of robotics in human genome research requires expertise in and interaction among a variety of disciplines, including molecular biology, engineering, and computing science. Hewlett-Packard, Inc., has provided the Human Genome Center at Lawrence Berkeley Laboratory (LBL) with a computer-driven robot for handling and processing biological samples. The robot consists of an active arm (a) capable of accurate and precise movement and of being programmed to change hands during procedures; eight pumps for dispensing and sampling very small volumes with comparable accuracy and precision (not in view); a spectrophotometer (b) for color analysis of the samples; rack towers and incubator hotels (c) that hold either unlidded plates (d) or racks of pipette tips (e); a hand tree (f) that holds tools for gripping (g) or pipetting with either a large single mandrill (h) or with 16 channels (i); a rake to scrape off used tips; and a blank hand for future customization. The control pole (j) is capable of five degrees of freedom: rotation, height, grip, reach, and wrist twist (disabled). The staff of the LBL divisions of Engineering, Computing Science, and Molecular Biology are working with the engineers of Hewlett-Packard to modify existing hardware, as well as to develop new software. Initial applications developed in this effort will speed the use of second- and third-generation robots in commercial, medical, and forensic laboratories. (Photograph provided by the Human Genome Center, Lawrence Berkeley Laboratory.)

Physical Mapping

Automated cosmid fingerprinting and contig assembly. Chromosome-19-specific cosmids are digested with restriction enzymes, and the fragments are labeled with fluorochromes. The Beckman Instruments, Inc., Biomek® robotic system processes sets of 48 cosmids per experiment. Throughput is increased, because of the capability to load the restriction fragments from three cosmids (each labeled with a different fluorochrome) plus size standards (a fourth fluorochrome) in each lane of a denaturing polyacrylamide gel. The labeled restriction fragments are detected, and the fluorescence is digitized as the fragments migrate past a laser beam in an Applied Biosystems 370 automated DNA sequencer. Fragment data acquisition may be monitored during electrophoresis to ensure that operation is proceeding normally.

In the large photograph, the fragment peaks from each of three cosmids are labeled in blue, green, and yellow. The size standard is in red. Several lanes of the gel are depicted on the top of this figure, and a historical view of the data from one lane is shown in the lower plot. The restriction fragment mobility data for each cosmid are analyzed by a suite of software programs developed at Lawrence Livermore National Laboratory. Fluorescence signals are processed to remove noise, to identify peaks (representing restriction fragments), and to calculate fragment lengths by comparison to the size standards.

The inset photograph demonstrates fragment size comparisons for all pairs of cosmids to determine whether they share a significant number of fragments of the same size (to within 1–1.5 bases). A single statistic is calculated that estimates the strength of the overlap. A best-overlap solution is determined and presented graphically for inspection, manipulation, and detailed query of the underlying database. The cosmids, represented by the warmer-color (red and yellow) bars, are those that exhibit the most overlap. (Photographs provided by Anthony Carrano, Lawrence Livermore National Laboratory.)


[Chn] 15]

Research Facility Narratives

Lawrence Berkeley Laboratory

Introduction

In September 1987, Lawrence Berkeley Laboratory (LBL) and Los Alamos National Laboratory (LANL) were designated as Human Genome Centers. LBL's response was to initiate an effort that would incorporate at one site all of the elements necessary for successful execution of the project and provide an environment in which integration of emerging concepts, methods, and techniques was immediate. Interdisciplinary efforts are a hallmark of LBL and of the other national laboratories. The unique aspect of the LBL Human Genome Center—the breadth of its activities—is made possible by the juxtaposition of the great variety of LBL talent in several areas (i.e., instrumentation, materials science, and computing technologies) with the large, outstanding biological research communities in the Berkeley and neighboring Bay Area institutions.

The Center's current activities are concentrated in four areas:

- construction of a physical map of the human genome,
- automation of existing physical mapping methods and development of new ones,
- enhancement of existing technologies for handling and sequencing DNA, and
- improvement of methods for interpreting and analyzing maps and sequence data.

Current efforts focus on chromosome 21—with 50 Mbp, the smallest human chromosome. Three principles guided the development of this research agenda. The first was the realization that new methods, techniques, and instrumentation must be developed to complete the genome project, and that the most effective way to do this would be to work in close physical and intellectual contact with pilot-scale mapping and sequencing efforts. The second principle was that new methods are more easily implemented on relatively small-scale projects. The third guiding principle was the projection that the program would be characterized by the use of newer and more powerful techniques for automated sample handling and biochemical analysis that would be needed for the increasingly larger data-producing projects.

Thus, development of improved data analysis and management methods was thought to be necessary both to handle the data generated at the LBL Center and to merge and reconcile these results with those from the many other laboratories involved in genome mapping and sequencing. The scientific direction of the Center is reviewed annually by an eight-member advisory committee, whose members include two Nobel laureates and six members of the U.S. National Academy of Sciences. The Center is involved directly with the University of California at Berkeley in a graduate training program in biotechnology; approximately half of this program's faculty are associated with the Center at LBL.

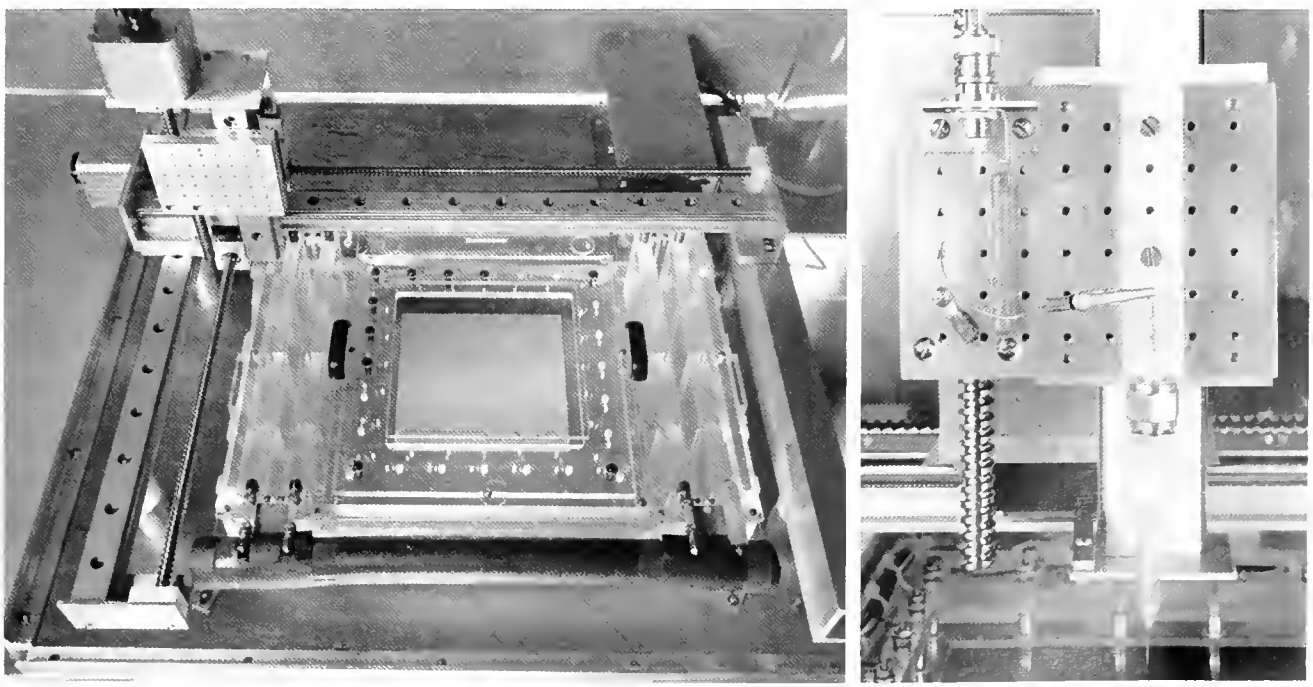
The unique features of LBL's current research program include the development of totally new DNA-handling procedures and physical mapping methods and the use of yeasts both as a source and as a testing ground of new techniques.

Accomplishments

Restriction maps of large regions of chromosome 21 completed. By using (a) single-copy probes with known regional locations to assign fragments and (b) partial digests, chromosome-specific *Not* I linking probes, and polymorphism among cell lines to assign neighboring bands, LBL researchers have identified about 40 Mb of DNA from chromosome 21 and linked *Not* I fragments in several regions of the *q* arm; these fragments include a continuous section that starts at the telomere and extends for 8.5 Mb.

Physical Mapping

A pulsed-field gel electrophoresis test bed. The basic DNA separation technique of modern molecular biology is gel electrophoresis—in particular, pulsed-field gel (PFG) electrophoresis. In the PFG test bed shown here (left), conditions at any point in the gel can be monitored by a probe mounted on a computer-controlled platform (right). Parameters such as current, gel temperature, and pH can now be monitored to optimize conditions and to assure reproducibility among different PFG sessions. The test bed allows active computer control of electrode potentials, as well as the capability for programming complex pulse cycles. The results of continuing studies with this and other test beds will be increased resolution and shorter separation times, especially for DNA fragments of more than 5 Mbp. (Photographs provided by the Human Genome Center, Lawrence Berkeley Laboratory.)



Research Facility Narratives: LBL

Restriction map of the entire *S. pombe* genome completed.

Second-generation mapping strategy designed—to be implemented using the polymerase chain reaction (PCR) and automated-sequencing procedures. LBL's strategy is to use DNA sequences from the ends of the fragments as unique "I. D." tags. Matching these DNA sequences with similarly sized DNA sequences from linking clones (clones that contain DNA from the ends of adjacent large fragments) will facilitate map construction. The advantages of this protocol over others is the speed with which results are generated, the precision of ordering, the simplicity of data analysis, and the fact that the mapping process generates sequence data as well.

Computer-controlled pulsed-field gel (PFG) electrophoresis apparatus constructed and used to discover new pulse shapes that speed separations five- to tenfold. The test apparatus consists of a 24-node computer-controlled power supply that is capable of independently programming the pulse sequence of individual electrodes. Computer calculations are used to define the electrode voltage distribution required to generate specific electric field profiles within the gel. During the run, a three-dimensional precision manipulator allows computer-controlled scanning of the gel to monitor parameters such as voltage, ionic strength, pH, and temperature.

Interactive computer processing of gel images developed so that collection of over- and underexposed areas eliminates the need for repeating exposure times. This program provides fully automatic analysis of the separated DNA, if desired; however, the operator has the option to substitute his/her own judgement at every step. The data sets of separated DNA are subsequently used for algorithmic comparisons between lanes and gels, as well as in mapping algorithms, and are stored in a laboratory database.

New computational algorithms for map assembly developed. Dynamic programming algorithms—traditionally, the technique of choice for matching problems—are likely to be impracticable for problems of the size generated by human genomic research. For this purpose, LBL has worked to extend the use of suffix trees; the objective is to develop computational methods that are faster and more practical than current ones.

New model for map assembly data management systems (based on extended-entity-relationship model) designed. LBL researchers are using a database schema design tool (SDT) developed at LBL to provide a powerful and easy-to-use interface for biologists and to increase the productivity of the database design process.

Mapping the telomere of human chromosome 4, which contains the gene for Huntington's disease, initiated. Probes have been developed to recognize a telomeric

area of chromosome 4. Various mapping techniques have been used to begin mapping around this area.

Method under development for testing single fluorescent molecules in flowing streams. Based on a new theory for optimizing fluorescence detection, an instrument has been developed for measuring single molecules of phycoerythrin. This single molecule detection (SMD) system can measure concentrations three orders of magnitude lower than conventional fluorescence detection systems. The SMD system is now being applied to the optimization of fluorescence detection of DNA fragments on sequencing gels.

Future Directions

- Develop methods for microsurgical dissection of single DNA molecules and PCR of fragments for direct mapping and sequencing.
- Implement a chromosome-21 database pilot project to facilitate cooperation among different groups and obtain user feedback on desirable or necessary features.
- Complete an ordered library of chromosome 21 by second-generation methods.
- Extend the size range of PFG to allow much higher resolution in the 1- to 10-Mbp range and separations of even larger DNAs.
- Increase the sensitivity of imaging techniques for chromosome in situ hybridization.
- Develop applications for robotics in laboratory procedures such as screening libraries for linking clones and PCR analysis of sliced gel lanes.
- Develop imaging techniques for direct sequence reading by scanning tunneling microscopy (STM) or atomic force microscopy (AFM).
- Incorporate direct-imaging plate technology into automated protocols.
- Develop efficient, versatile algorithms for searching and matching strings in databases.
- Extend data thesaurus techniques of consistent naming and prototypes to store, index, search, and retrieve genetic map information.
- Automate genetic stock centers to conduct pilot studies for acquisition, evaluation, maintenance, and distribution of clones and associated data.

For more information on the LBL Human Genome Center, please contact Charles R. Cantor, Director, at (415) 486-6800 or FTS 451-6800.

Lawrence Livermore National Laboratory

Research Facility Narratives

Introduction

The human genome project at Lawrence Livermore National Laboratory (LLNL) is a multidisciplinary effort. It draws upon the Livermore matrix organization to bring together a team of chemists, molecular biologists, physicists, mathematicians, computer scientists, and engineers in an interactive research environment. The broad goals of the project are to:

- develop biological and physical resources useful for genome research,
- model and evaluate DNA mapping and sequencing strategies,
- combine these resources and strategies in an optimal way to construct ordered-clone maps and DNA sequences of human chromosomes, and
- use the map and sequence information to study genome organization and variation.

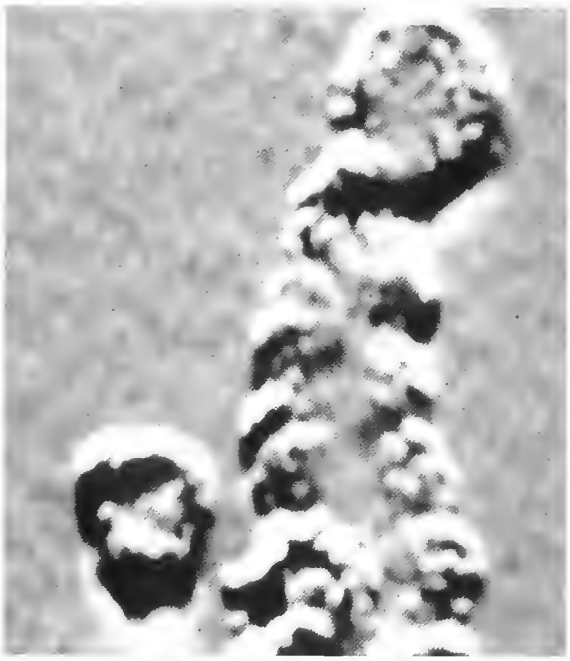
Livermore's entry into genomics research was facilitated by existing scientific interest, expertise, and research in molecular biology, cytogenetics, mutagenesis, and instrumentation development, as well as by participation in the National Laboratory Gene Library Project. In addition, the programs at Livermore have contributed substantially to the identification and characterization of human DNA repair genes and specifically to the three DNA repair genes on chromosome 19. It is not unexpected then that LLNL's initial interest focused on this particular chromosome. Because Livermore's program is multidisciplinary, a variety of scientific talent can be drawn upon to meet its needs. Livermore's role in the DOE Human Genome Program is one of technology development, validation, and application to ongoing and new programs in structural biology and mutagenesis. The present human genome effort at Livermore involves several interactive research components that have as a common goal the construction of ordered-clone maps of the human genome. These component tasks include:

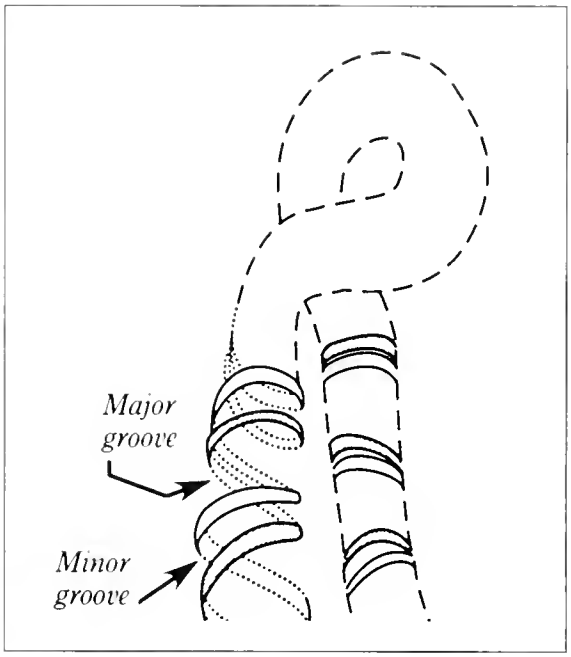
The National Laboratory Gene Library Project. This project is a joint effort with Los Alamos National Laboratory and has as its goal the construction of human-chromosome-specific libraries in lambda, cosmid, and yeast artificial chromosome (YAC) vectors for use in physical mapping and other studies. The project draws upon experience in flow instrumentation and chromosome sorting at these two national laboratories.

Resource Development and Management. The group working on this project is responsible for several tasks, including the construction of specialized cloning vectors and recombinant libraries for application to physical map construction; the development of new biochemical and biophysical techniques for mapping and sequencing; and the management and distribution of Livermore's material resources such as cell lines, probes, library arrays, filters, and DNA.

Mathematics and Computations. This project's staff is responsible for mathematical modeling of mapping strategies, the development of data analysis algorithms, data processing, and the construction and maintenance of interactive relational databases for internal and external access.

Map Construction. Using material resources, techniques, and computational methods, the group assigned to this task is assembling ordered-clone maps. This task involves the application of methods for clone fingerprinting, large-fragment electrophoresis, multiplex walking strategies, and in situ hybridization to chromosomes.





Sequencing Technologies

Imaging by scanning tunneling microscopy (STM). In a collaborative effort, Lawrence Livermore National Laboratory and Lawrence Berkeley Laboratory scientists have used STM to image native, unstained DNA molecules under conditions of normal atmospheric pressure. The image here, as illustrated in the schematic depiction beside it, is sufficiently resolved to show the major and minor grooves of the DNA double helix. Some researchers have even suggested that, with further development, STM imaging may someday be an important technology used to sequence DNA. [Photograph first published by the American Association for the Advancement of Science in Thomas Beebe, Jr., et al., "Direct Observation of Native DNA Structures with the Scanning Tunneling Microscope," *Science* **243**, 370–372 (Jan. 20, 1989). Photograph and figure provided by the Human Genome Center, Lawrence Berkeley Laboratory.]

Instrumentation and Automation. The group responsible for this task develops instrumentation that will facilitate the handling of DNA and cloned materials so that human involvement in highly repetitive tasks can be minimized.

High-Resolution Imaging. In this project, two novel approaches to DNA sequencing are being developed using scanning tunneling microscopy (STM) and high-resolution imaging with X-ray lasers.

The research groups involved in these component tasks are highly interactive. Individual staff members often have responsibilities in more than one group. Collaborating with other research groups throughout the world, Livermore coordinates its research activities with others who are either involved directly in the human genome effort or who have mutual scientific interests.

Accomplishments

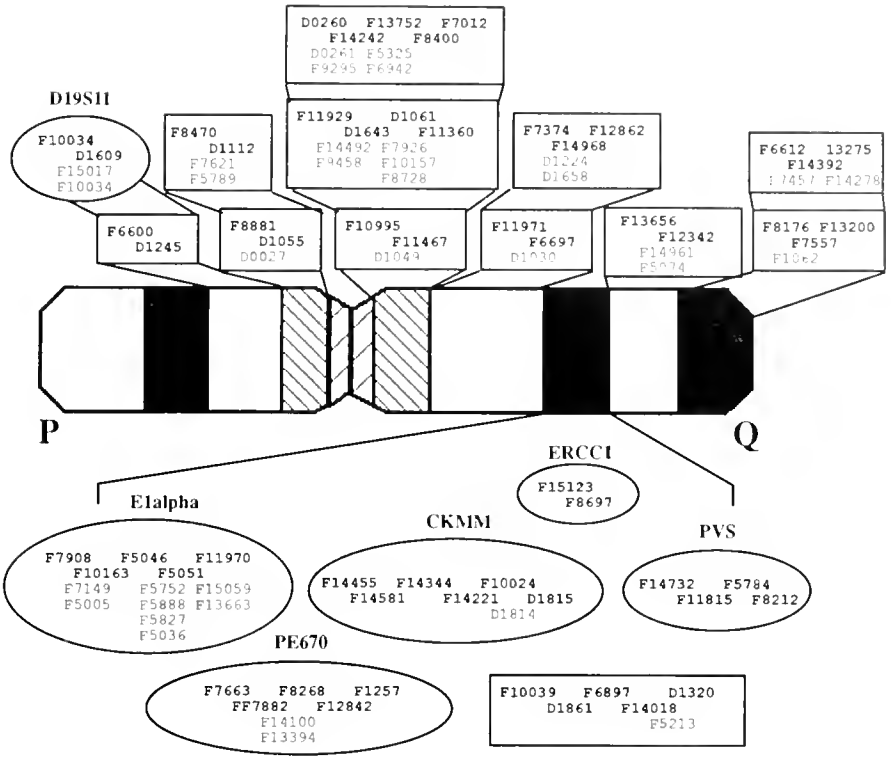
In the past two years, Livermore has made excellent progress in the construction of chromosome-specific libraries, in the development and application of new biochemical and mathematical approaches for constructing ordered-clone maps, in the automation of fingerprinting chemistries, and in high-resolution imaging of DNA.

As part of the National Laboratory Gene Library Project, small insert libraries of each human chromosome were constructed using material purified by flow sorting. These libraries have been deposited with the American Type Culture Collection for worldwide distribution. Construction of large-insert partial-digest libraries in both lambda and cosmid vectors has begun; currently, five such libraries are being characterized for chromosomes 11, 19, 21, 22, and Y.

In the area of resource development, new or modified existing vectors have been constructed to:

- facilitate cloning small amounts of DNA in cosmids,
- clone *Not* I linking probes in lambda and plasmids, and
- clone large fragments of DNA as YACs.

Several of the cosmid vectors have been transferred to the U.S. commercial sector. They were used to construct cosmid libraries of flow-sorted chromosomes. The plasmid and lambda vectors were used to create a small *Not* I linking library of chromosome 19. A library of chromosome 19 in YACs is currently being expanded. A probe repository for chromosome 19 has been established, and the collection is growing; these probes have been used to screen Livermore's chromosome-19 libraries. Individual cosmid clones and filters containing DNA from individual clones of chromosome 19 are being distributed to the scientific community as part of a collaborative effort to map this chromosome.



Physical Mapping

Chromosome-19 contig map as of October 1989. Of the 2200 cosmids processed for fingerprint analysis, over 900 of them are in contigs with an average contig length of about 3 cosmids. Selected cosmid contigs are shown in this figure with their map location as determined by fluorescence in situ hybridization. Each cosmid is designated by its clone number; bold numbers represent the "minimal" covering set for that contig. Contigs associated with known genes are in ovals, and the gene designation is indicated above each oval. The graph at the bottom of the figure shows the rate of contig formation. (Figure provided by Anthony Carrano, Lawrence Livermore National Laboratory.)

Research Facility Narratives: LLNL

To construct a set of cosmid contigs (overlapping clones) for chromosome 19, LLNL has developed a semiautomated fluorescence-based strategy for fingerprinting each clone. For this procedure, researchers use a robotic system to attach fluorophores to the ends of restriction fragments from each cosmid clone. Fragment lengths are determined using a commercially available laser scanning device to acquire data in real time from a polyacrylamide gel. The data collection is multiplexed in the sense that up to four different fluorophores (i.e., four clones) can be run in each gel lane. In this configuration, 48 cosmids can be analyzed per gel run.

Livermore has developed software that will process the acquired fingerprint signals, convert the information to a fragment length for each cosmid, use the fragment length data to compute a statistical measure of overlap between cosmids, and allow users to display data graphically and browse among the cosmid contigs. The fingerprinting procedure, from fluorochrome labeling through data processing and contig visualization, is now automated.

About three thousand cosmids have been processed to date. Researchers at LLNL have established over 330 cosmid contigs for chromosome 19—estimated to span 30% of the chromosome or about 20 Mbp. In addition, they established 6 cosmid contigs that span approximately 600 kb of chromosome 14. Several of the chromosome-19 contigs represent known gene loci; the others are located throughout the chromosome. A number of these contigs have been validated by restriction fragment digests or by in situ hybridization to metaphase chromosomes.

In a study parallel to contig map construction, Livermore is developing a large-fragment restriction map of chromosome 19; this approach will eventually fill the gaps between contigs. LLNL researchers have discovered that two of the DNA repair genes on this chromosome lie within 260 kb of each other. In addition, they have devised a technique to isolate region-specific probes based upon polymerase chain reaction (PCR) amplification of DNA located between human *Alu*-repetitive sequences. These probes are being used to identify those cosmids (from Livermore's chromosome-19 library) that span a specific region of the chromosome.

In collaboration, Lawrence Berkeley Laboratory and LLNL have used scanning tunneling microscopy to produce the highest resolution image of DNA to date. Further work at Livermore is focusing on the application of this technique to DNA sequencing. Advances have been made in DNA deposition techniques, construction of a new computer-controlled scanning tunneling microscope, refinement of the electronics to minimize noise, and development of new computer programs to improve visualization and analysis of imaged DNA.

Future Directions

- Livermore's highest priority is to complete, to the extent possible, an ordered-clone map of chromosome 19. This map will likely be a composite linear array of cosmid, lambda, and YAC clones. Other goals are to correlate this physical map with the genetic map and assist the scientific community in the localization and isolation of all the genes from chromosome 19. Livermore researchers will use state-of-the-art sequencing technology to sequence selected high-interest regions of the chromosome. Once the technology for map construction for a large portion of chromosome 19 has been validated, they will scale their efforts to other chromosomes.

- At some point in the human genome project, emphasis will shift from mapping to sequencing. Livermore plans to use large fragments such as cosmids or YACs as templates to explore rapid DNA sequencing methods that can be automated. The STM and X-ray imaging technologies at Livermore will be utilized in the sequencing effort, if appropriate.

- Because automation is an essential element of the physical mapping process, Livermore will continue to explore new processes and instruments to reduce the need for human involvement in the highly repetitive tasks. For example, LLNL will complete the development of a prototype cosmid DNA extractor and evaluate its utility. A number of other instruments for clone manipulation and biochemical processes will also be considered for automation.

- To assist in the completion of the ordered clone maps, Livermore's interaction with the scientific community is critical and will continue to be given high priority. An LLNL facility will serve as a resource laboratory for clones and for map information on chromosomes of interest. Ultimately, map and sequence information developed at Livermore will be used to study the global architecture of the chromosome and to evaluate somatic and genetic variation, spontaneous and induced, in man.

For more information on human genome research at LLNL, please contact Anthony V. Carrano at (415) 422-5698 or FTS 532-5698.

Los Alamos National Laboratory

Research Facility Narratives

Introduction

In several of the key technologies necessary for the DOE Human Genome Program, Los Alamos National Laboratory's (LANL's) experience and capabilities include:

- a strong core of molecular biology expertise,
- flow cytometry for sorting chromosomes and for single molecule detection,
- organization of databases (GenBank®), and
- computer analysis of nucleic acid sequences.

LANL has built upon this experience to develop resources, technologies, and strategies for mapping and sequencing the human genome and for organizing and analyzing the resulting data.

Some of the activities of the Human Genome Program require centralization and close coordination. Among these are (1) the provision of arrayed cosmid or yeast artificial chromosome (YAC) libraries from sorted chromosomes that will serve as reference material for physical mapping and (2) the assembly of the physical mapping and sequencing information into a computer database. LANL is continuing to play a leading role in providing these critical resources and in utilizing the findings and materials from the genome program for basic research.

LANL will collaborate with private industry, both to utilize the skills and resources of the private sector for strengthening the Los Alamos genome program and to assure the effective transfer of technologies to the U.S. commercial sector.

Accomplishments

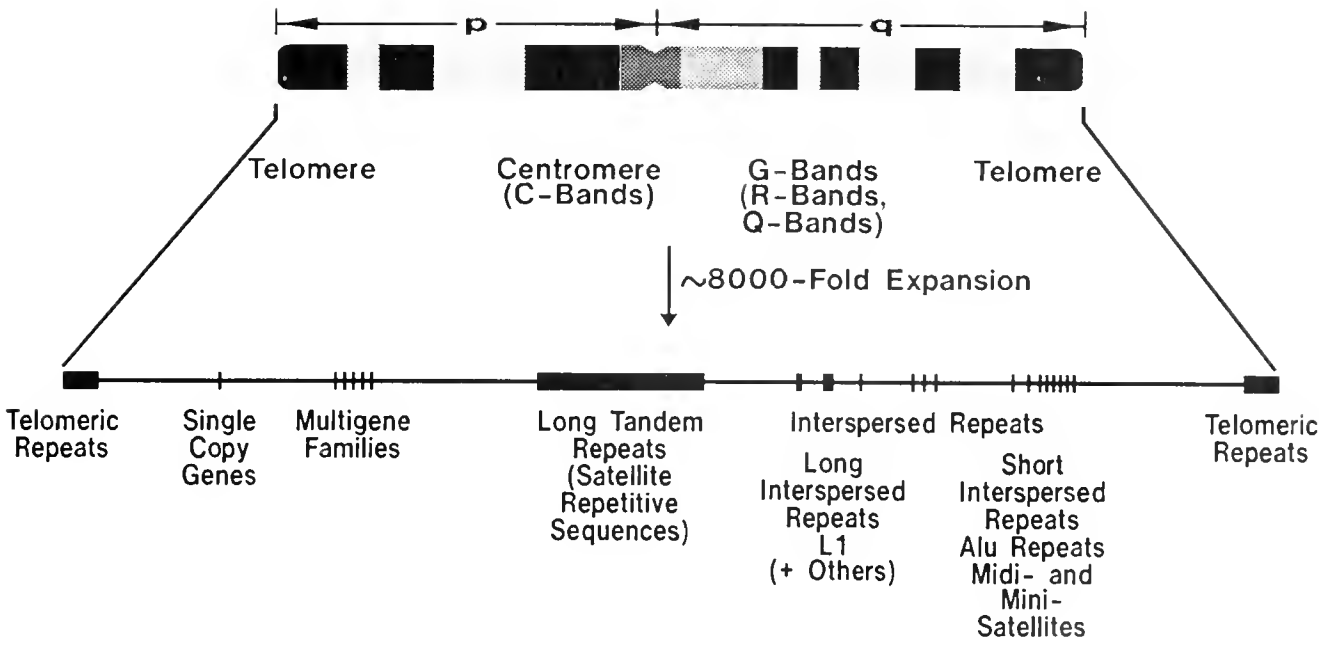
LANL investigators identified and cloned the human telomere sequence, TTAGGG, repeated several hundred times at the end of each human chromosome. Significant features of this discovery for the Human Genome Program include the following:

- The sequence provides definitive ends to the map of each chromosome and useful starting points for physical maps.

- The identification of the human telomere allowed 100,000–200,000 nucleotide human telomeric DNA sequences to be cloned in YAC vectors (in collaboration with M. Olson, Washington University). These sequences have been used to construct contig maps of the ends of several human chromosomes.

Physical mapping of human chromosome 16 is well under way at LANL, where a new approach has been developed to identify overlapping cosmid clones by exploiting the

high density of repetitive sequences in complex genomes. Individual clones are fingerprinted using a combination of restriction enzyme digestions followed by hybridization with selected classes of repetitive sequences. Along with information on the lengths of all restriction fragments, the occurrence of repeat sequences is acquired by image capture. With this fingerprinting data, overlapping cosmid clones are identified. Cosmid clones obtained from flow-sorted chromosome 16 were used to identify 2261 individual clones arrayed in 389 contig sets—approximately one-half of chromosome 16. The approach of "nucleating" at specific regions in the human genome and exploiting the high density of interspersed repetitive sequences in human DNA allows (1) rapid progress in early contig mapping phases that have generated large (>100 kb) contigs and (2) the production of a contig map with landmarks useful for rapid integration of the genetic and physical maps.



Physical Mapping

Organization of human chromosomes. This illustration summarizes the major types of sequences and regions that have been characterized on a human genome—or mammalian genome, in general. (Figure provided by C. E. Hildebrand, Los Alamos National Laboratory.)

Research Facility Narratives: LANL

In the National Laboratory Gene Library Project, libraries using the vector Charon 40 and cosmid libraries using the vector sCos 1 have been constructed for human chromosomes 16, 5, 8, and 4. A lambda library has been made for the X chromosome.

Progress has been made in the detection of single fluorescent molecules in a flowing liquid—an essential step in LANL's proposed system for sequencing single DNA molecules at a rate of ~10^3 bp/s. In this approach, the molecule is excited by a laser. LANL has markedly enhanced the signal-to-noise ratio so that single molecules such as fluorescein may be detected reliably.

A physical mapping database pilot project has been designed and is being used to manage data accumulating in the physical mapping project at LANL. A relational database has been established and is being managed with the Sybase data management system. Every clone is given a unique identifier and an arbitrary number of characteristics such as source, restriction fragments derived by various digests, restriction map, probe hybridization, relation to other clones, and relation to genetic markers or sequences.

A process has been established for identifying industrial partners. A workshop, attended by 24 companies, was held at Los Alamos, and proposals from 8 of those companies that responded to a request for proposal (RFP) are now under review.

Future Directions

- Establish, jointly with Lawrence Livermore National Laboratory (LLNL), a resource to make available arrayed libraries of cosmid clones for all the human chromosomes. The generation of YAC libraries from flow-sorted material will be investigated.
- Continue physical mapping of chromosome 16 with cosmid clones. Strategies and techniques for linking cosmid contigs will be developed, mostly with YAC clones; mapping of additional chromosomes will be initiated. Clones will be distributed to provide ties between physical and genetic linkage maps.
- Establish an integrated pilot program for sequencing of megabase regions generated by physical mapping.
- Develop a system for sequencing single DNA molecules at a rate of ~10^3 bp/s.
- Develop computational tools to support the Library Resource, clone characterization for physical mapping, and assembly of the physical map from clone overlap probabilities.
- Develop an integrated database for physical mapping and sequence information (linked to the genetic mapping database) plus computation, communication, and analysis tools to make them accessible at scientific workstations in molecular biology laboratories.

- Investigate the structure and function of repetitive sequences in the human genome.
- Study the organization and function of chromatin.
- Develop analysis programs for detecting and characterizing functionally significant patterns in genomic DNA.
- Collaborate with private companies to ensure the effective transfer of technology to the commercial sector.

For more information on the LANL Center for Human Genome Studies, please contact Robert K. Moyzis at (505) 667-3912 or FTS 843-3912.

Physical Mapping

Identification and cloning of the human telomere to define the ends of the human genetic and physical maps. Telomeres are defined as the ends of chromosomes. These specialized structures are involved in the replication and stability of linear DNA molecules. Investigators at Los Alamos National Laboratory (LANL) have identified and cloned the human telomere [Moyzis et al., *Proc. Natl. Acad. Sci. USA* **85**, 6622–6626 (1988)]. Fluorescence in situ hybridization has been used, in addition to other biochemical and biophysical techniques, to localize this unusual sequence, $(TTAGGG)_n$, to human telomeres. Seen in the inset photograph as fluorescent yellow spots on red-stained human chromosomes, this sequence is present at the telomeres of all vertebrate species and, hence, must have arisen over 400 million years ago [Meyne et al., *Proc. Natl. Acad. Sci. USA* **86**, 7049–7053 (1989)].

The ultimate proof that this repeating DNA sequence TTAGGG is the human telomere is to show that the sequence functions as a unit in an artificial chromosome. In collaboration with the staff of Maynard Olson's laboratory at Washington University, the LANL staff was able to construct yeast artificial chromosomes (YACs) in which the natural human $(TTAGGG)_n$ sequence functioned as a telomere in yeast cells [Riethman et al., *Proc. Natl. Acad. Sci.* 86, 6240–6244 (1989)]. These results indicate that the yeast telomere replication machinery can indeed recognize the human telomere, even though the common ancestor of yeast and humans lived over one billion years ago.

In addition to demonstrating that the $(TTAGGG)_n$ sequence functions as a telomere, these YACs allowed large (100,000–200,000) nucleotide fragments to be isolated from the ends of human chromosomes. Seen in the large photograph is in situ hybridization of DNA from one of these YAC clones that originated from the telomere of human chromosome arm 7q. DNA from such YACs can be used to define the ends of the human genome genetic and physical maps.

LANL's discovery of the human telomere is a significant milestone in efforts to map the human genome. Prior to identifying the telomeric sequence, investigators were without a reference point from which to orient DNA mapping studies for identification of DNA markers that would be useful for the analysis of disease genes. [The inset photograph was first published in *Proc. Natl. Acad. Sci. USA* **85**, 6622–6626 (1988). The large photograph was first published in *Proc. Natl. Acad. Sci. USA* **86**, 6240–6244 (1989). Photographs provided by Robert Moyzis, Los Alamos National Laboratory.]

Abstracts of DOE-Funded Research

The abstracts in this section were contributed by the DOE Human Genome Program grantees and contractors. The names of the principal investigators are in bold; the address and phone number following each abstract title are those of the principal investigator. If more than one principal investigator is listed with an abstract, the address and phone number belong to the first. An index of project categories and principal investigators is given at the beginning of this section. Listed at the end is an index of all project investigators named in the abstracts.

Project Categories and Principal Investigators

Abstracts

Principal investigators of the research projects described by the abstracts in this section are listed here under their respective subject categories.

Resource Development

Physical Mapping

Mapping Instrumentation

Sequencing Technologies

Informatics

Small Business Innovative Research (SBIR)— Phase I (1989 Awards)

Small Business Innovative Research (SBIR)— Phase II (1988, 1989 Awards)

Index to Project Investigators

Abstracts

Monochromosomal Hybrids for the Analysis of the Human Genome

Raghbir S. Athwal
Department of Microbiology and Molecular Genetics, University of Medicine and Dentistry of New Jersey, New Jersey Medical School, Newark, NJ 07103-2757
(201) 456-4484

In this research project we have proposed to develop rodent/human hybrid cell lines, each containing a single different human chromosome. The human chromosomes will be marked with *Eco*gpt and stably maintained by selection in the hybrid cells.

This experimental approach to producing the proposed cell lines involves the following: Using a retroviral vector, we will first transfer a cloned selectable marker, *Eco*gpt (an *E. coli* gene for xanthine-guanine phosphoribosyltransferase: XGPRT), to normal diploid human cells. The transferred gene will integrate at random into multiple sites in the recipient cell genome. Clonal cell lines from independent transgenotes will each carry the selectable marker integrated into a different site and perhaps a different chromosome. The chromosome carrying the selectable marker will then be transferred further to mouse cells by microcell fusion. In addition, we will use directed integration of *Eco*gpt into the chromosome present in rodent cells, not otherwise marked with a selectable marker. This will allow us to complete the bank of proposed cell lines.

Since the human chromosome will be marked with a selectable marker, it can be transferred to any other cell line of interest for complementation analysis! Clones of each cell line, containing varying sized segments of the same chromosome produced by selection for the retention or loss of the selectable marker following X-irradiation or by metaphase chromosome transfer, will facilitate physical mapping and determination of gene order on a chromosome.

Human Recombinant DNA Library

Larry L. Deaven, Robert K. Moyzis, Jon Longmire, and C. E. Hildebrand
Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-3114, FTS 843-3114

The goal of the National Laboratory Gene Library Project (NLGLP) is the production of chromosome-specific human gene libraries and their distribution to the scientific community (1) for studies of the molecular biology of genes and chromosomes, (2) for the study and diagnosis of genetic disease, and (3) for the physical mapping (ordering) of chromosomes. This is a cooperative project employing the flow-sorting and molecular-cloning expertise at the Los Alamos National Laboratory (LANL) and the Lawrence Livermore National Laboratory. The specific aim of Phase I of the project was the production of complete digest libraries from each of the human chromosomal types purified by flow sorting; the average insert size expected was about 4 kb. The bacteriophage lambda vector was Charon 21A, which has both *Eco*R I and *Hin*d III insertion sites accommodating human DNA fragments 0–9.1 kb in size. Each laboratory has produced a complete set of chromosome-specific libraries: LANL with *Eco*R I and LLNL with *Hin*d III. The small insert libraries are deposited in a repository at the American Type Culture Collection, Rockville, Maryland; over 2000 aliquots have been distributed to over 500 laboratories worldwide.

The second phase of the project—the construction of partial digest libraries with larger inserts in more advanced, recently developed lambda vectors (9–23 kb) and in cosmid vectors (33–46 kb)—is under way. These large insert libraries have characteristics that are better suited to basic studies of gene structure and function, organization of genes on chromosomes, and ordering of cloned sequences. The Phase II strategy is to split the genome between the two laboratories, with Livermore cloning 12 chromosomal types (starting with 7, 11, 19, 21, 22, and Y) and Los Alamos cloning the other 12 (starting with 4, 5, 8, 16, 17, and X). In this way, each chromosomal type will be cloned into both lambda and cosmid vectors. Vectors currently in use include Charon 40 and lambda GEMII (phage) and sCos1 and Lawrist 5 (cosmid). Partial digest libraries have been constructed in either phage or cosmid vectors for chromosomes X, Y, 16, 19, 21, and 22.

Field-Flow Fractionation of Chromosomes

J. Calvin Giddings
Department of Chemistry, University of Utah, Salt Lake City, UT 84112
(801) 581-6683

Field-flow fractionation (FFF), a powerful and versatile methodology of relatively recent origins, is applicable to the separation of virtually all categories of macromolecules and particles. The object of this study is to apply state-of-the-art field-flow fractionation methods to chromosomes, in an effort to separate and purify them from one another and from background cellular debris. This research is focused primarily on the utilization of sedimentation/steric FFF for this problem, but other FFF techniques, including flow FFF, may be involved as well. Recent experiments involving particles in the size range of chromosomes demonstrate the feasibility of working in the chromosome-size range. In all likelihood, FFF methods have sufficient flexibility to circumvent any potential problems encountered in chromosome separation, such as chromosome adsorption or disruption.

New Dyes for DNA Sequencing

Hee-Chol Kang, James E. Whitaker, Peter C. Hewitt, and **Richard P. Haugland**
Molecular Probes, Inc., Eugene, OR 97402
(415) 486-5717

We have been actively synthesizing and evaluating sets of new fluorophores that can be excited by the argon laser at 488 or 514 nm for possible use in DNA sequencing. The objectives are to synthesize sets of four dyes whose emission spectra have relatively low overlap, whose fluorescence when excited with the argon laser is brighter than currently available fluorophores, and whose properties of ionic charge are uniform for minimum interference with electrophoretic separations. Principal among the dyes prepared have been fluorescein-rhodamine bifluorophores in which the energy absorbed by the fluorescein is emitted almost totally at the rhodamine emission wavelength. Examples of these dyes have been prepared where the energy transfer has been >98% efficient with pseudo-Stokes shifts of up to 100 nm. Several reactive versions of rhodamine and of rhodol dyes have been prepared which fluoresce when excited by the argon laser and whose emission is brighter than tetramethylrhodamine. The fourth class of new fluorophores with potential for use in DNA sequencing are reactive, boron dipyrromethene difluoride (Bodipy™) derivatives, which have been prepared in several reactive forms. Probes derived from this fluorophore have unusually narrow emission band width and have high absorbance and quantum yield. The prospects for preparation of new DNA sequencing dyes with higher detectability and spectral resolution will be presented.

Optimizing Procedures for a Human Genome Repository

William C. Nierman and **Donna R. Maglott**
American Type Culture Collection, Rockville, MD 20852
(310) 231-5559

The cloned genes and DNA fragments identified during the human genome project should be stored in a repository and made available to the research community. Such a repository would also establish a set of reference clones to facilitate comparison of data generated from different laboratories.

Repositories of well-characterized cloned human DNA fragments currently exist, but at a much smaller scale than necessary for the human genome project. Procedures used in these repositories cannot be expanded without modification. Methods must be improved to automate DNA preparation; clone verification; data maintenance and analysis; and sample storage, recovery, and distribution. Procedures reducing the amount of sample needed for verification and storage must be perfected. The objective of this project is to establish a pilot repository to evaluate such protocols and instrumentation. Initial emphasis will be placed on automating clone verification by analyzing restriction fragments on a DNA sequencing machine and comparing fragment sizes to those already obtained by depositors. Methods will also be explored to use robotics for DNA preparation; to manage information effectively; to verify clones for which there is no restriction data; and to improve methods of sample storage, retrieval, and distribution. These procedures will be tested through the development and operation of a pilot repository using the contigs of lambda clones identified by Maynard Olson's laboratory for the *S. cerevisiae* genome, and chromosome-16- and chromosome-19-specific contigs identified by the Los Alamos and Lawrence Livermore national laboratories.

DNA Sequence Analysis with Modified Bacteriophage T7 DNA Polymerase

Stanley Tabor, Hans E. Huber, John Rush, and **Charles C. Richardson**
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115
(617) 732-1864

The 3′ to 5′ exonuclease activity of phage T7 DNA polymerase (gene 5 protein) can be inactivated selectively by reactive oxygen species. The chemically modified enzyme is highly processive in the presence of *E. coli* thioredoxin and discriminates against dideoxynucleoside triphosphates (ddNTPs) only four- to sixfold. Consequently, dideoxynucleotide-terminated fragments have highly uniform radioactive intensity throughout the range of a few to thousands of nucleotides in length. There is virtually no background due to terminations at pause sites or secondary-structure impediments in the template. Chemically modified gene 5 protein, by virtue of having low exonuclease activity, has enzymatic properties that distinguish it from native gene 5 protein. We have exploited these properties to show by a chemical screen that modification of a histidine residue reduces selectively the exonuclease activity. In vitro mutagenesis of histidine 123, and of the neighboring residues, results in varying reduction of the exonuclease activity. A deletion of 28 amino acids that encompasses His123 eliminates all exonuclease activity (< 10^{-6} %). Incorporation of ddNTPs by T7 DNA polymerase and *E. coli* DNA polymerase I is more efficient when Mn^{2+} rather than Mg^{2+} is used for catalysis. Substituting Mn^{2+} for Mg^{2+} reduces the discrimination against ddNTPs approximately 100-fold for DNA polymerase I and 4-fold for T7 DNA polymerase. With T7 DNA polymerase and Mn^{2+}, ddNMPs and dNMPs are incorporated at virtually the same rate. Mn^{2+} also reduces the discrimination against other analogs with modifications in the furanose moiety, the base, and the phosphate linkage. The lack of discrimination against ddNTPs using the genetically modified T7 DNA polymerase and Mn^{2+} results in uniform terminations of DNA sequencing reactions, with the intensity of adjacent bands on polyacrylamide gels varying in most instances by less than 10%. A novel procedure that exploits the high uniformity of bands can be used for automated DNA sequencing. A single reaction with a single labeled primer is carried out using four different ratios of ddNTPs to dNTPs; after gel electrophoresis in a single lane, the sequence at each position is determined by the relative intensity of each band.

For more information see the following articles by S. Tabor and C. C. Richardson: *Proc. Natl. Acad. Sci. USA* **84**, 4767–4771 (1987), *J. Biol. Chem.* **264**, 6447–6458 (1989), and *Proc. Natl. Acad. Sci. USA* **86**, 4076–4080 (1989).

Human Repetitive DNA Sequences for Use as Markers in Mapping the Human Genome

Esther P. Leeflang, Gui-Lin Wang, Joe M. Gatewood, and **Carl W. Schmid**
Department of Chemistry, University of California, Davis, CA 95616
(916) 752-3003

A library of repetitive human DNA sequences was constructed from renatured DNA and subsequently screened with known repeats. Of the 460 clones examined, 267 did not hybridize with any of the known repetitive DNAs. Following preliminary sequence analysis and copy-number determination, ten of the clones were selected for further study. The repetitive DNA clones were used as hybridization probes to isolate lambda phage clones from a genomic library. The base sequence, copy number, genomic arrangement, and evolutionary divergence of the new repeat families are now being analyzed.

Gene Libraries for Each Human Chromosome: Construction and Distribution

Marvin A. Van Dilla, Pieter de Jong, Barbara Trask, Anthony V. Carrano, Joe Gray, Kathy Yokobata, and Ger J. van den Engh
Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
(415) 422-5662, FTS 532-5662

The goal of the National Laboratory Gene Library Project (NLGLP) is the production of chromosome-specific human gene libraries and their distribution to the scientific community for the diagnosis and study of genetic disease, determination of the structure and function of genes, and for the physical mapping of chromosomes. This cooperative project employs the flow-sorting and molecular-cloning expertise at the Los Alamos and Lawrence Livermore national laboratories. The specific aim of Phase I of the project was the production of complete digest libraries from each of the human chromosomal types purified by flow sorting. The bacteriophage lambda vector used was Charon 21A, which has both *Eco*R I and *Hin*d III insertion sites accommodating human DNA fragments 0–9.1 kb in size. Each laboratory has produced a complete set of chromosome-specific libraries, LANL with *Eco*R I and LLNL with *Hin*d III. Library purity ranges from nearly 100% for good chromosome preparations and favorably placed peaks in the flow karyotype to about 50% for some early preparations from cell lines with unfavorably placed peaks. The libraries are deposited in a repository at the American Type Culture Collection (ATCC), Rockville, Maryland; about 2400 aliquots have been distributed to over 500 laboratories worldwide. All Livermore libraries have been subcloned into the plasmid vector Bluescribe (Stratagene, La Jolla, California), facilitating both the use of the DNA probes and the preparation of RNA end probes.

Phase II, the construction of libraries with large inserts in lambda replacement vectors (accept about 9–23 kb) and in cosmid vectors (accept about 33–46 kb), is under way. These large insert libraries have characteristics that are better suited to basic studies of gene structure and function, organization of genes on chromosomes, and ordering of cloned sequences. The Phase II strategy is to split the genome cloning responsibility between the two laboratories [i.e., Livermore will clone 12 chromosomal types (1, 2, 3, 7, 9, 11, 12, 18, 19, 21, 22, and Y), and Los Alamos will clone the other 12 (4, 5, 6, 8, 10, 13, 14, 15, 16, 17, 20, and X)]. In this way, each chromosomal type will be cloned into both lambda and cosmid vectors. Livermore is using the lambda vector Charon 40 (accepts 10–25 kb inserts) and, more recently, lambda GEM11, which has about the same acceptance range as Charon 40 but is particularly suited for the manipulations required to efficiently clone, map, sequence, and "walk" along contiguous segments of genomic DNA. At Livermore, the cosmid vector is Lawrist 5 (accepts inserts of 34–46 kb), which has the same advantageous features for users as lambda GEM11 and is double the insert size. The cloning procedures (more complicated than for Phase I) have now been worked out, and we have constructed two large Charon 40 libraries for

chromosome 19; large lambda GEM11 libraries for chromosomes 11, 21, 22, and Y; and Lawrist 5 libraries for all these chromosomal types except 11. Phase II libraries are characterized as fully as resources allow both in-house and by a small number of interested, high-quality external laboratories before release to ATCC. Currently, this characterization by test labs is in progress for all but one of these libraries. The one exception is the chromosome-19 library in Charon 40; positive characterization results led to the release of this library to the ATCC in August 1988.

New Approaches for Constructing Expression Maps of Complex Genomes

Sherman M. Weissman, R. Kandpal, A. Swaroop, S. Parimoo, H. Arenstorf, and D. Ward
Department of Human Genetics, Yale University, New Haven, CT 06510
(203) 785-2677

The overall objective of this project is to develop and demonstrate methods for preparing normalized cDNA libraries and using them for gene mapping and mutation detection. A phage vector has been prepared that can be used efficiently to generate single-stranded cDNA clones. A number of human sources have been used to prepare the cDNA libraries. Biotin avidin selection methods have been developed for convenient preparation of subtracted cDNA libraries and are currently being evaluated for their ability to generate selected cDNA libraries containing only those cDNA complementary to selected segments of the human genome. In addition, polymerase chain reaction methodology is being adapted to make chromosome jumping a much more efficient general procedure for long-range genome mapping and to provide improved methods for preparing selected cDNA libraries.

Abstracts

Human Chromosome 21: Linkage Mapping and Cloning DNA in Yeast Artificial Chromosomes

S. E. Antonarakis, P. A. Hieter, and M. K. McCormick
Center for Medical Genetics, The Johns Hopkins University School of Medicine, Baltimore, MD 21218-2608
(301) 955-7872

The goal of our research is to contribute to the cloning of human chromosome 21 DNA in yeast artificial chromosomes (YACs). Chromosome 21 is the smallest human chromosome and contains about 1.4% of the human genome. The cloning of human DNA in YACs (Burke et al., *Science* **236**, 806–812, 1987) allows large fragments of DNA (100–1000 kb) to exist as additional chromosomes in *S. cerevisiae*. We used new YAC cloning vectors that facilitate the manipulation and mapping of the resulting YACs. DNA from cell line WA 17 (a mouse-human hybrid with chromosome 21 as the only human material) and from flow-sorted chromosome 21 were used as the starting material. Size-selected DNA from complete *Not* I or partial *Eco*R I digestion was ligated to the vectors, and yeast spheroplasts were transformed in the presence of polyamines to eliminate a bias in favor of smaller DNA inserts. In our initial experiments, YACs have been obtained from both DNA sources; the average size of those from the WA-17 cell line was 410 kb.

Specific future research goals include mapping the YAC clones and scaling up the experiments in order to obtain a large number of YACs, linking the YACs in overlapping contigs, and constructing a macrorestriction map of chromosome 21.

Molecular Mapping of Chromosomes 17 and X

David F. Barker, Huntington F. Willard, Pamela R. Fain, Arnold R. Oliphant, and David E. Goldgar
Department of Genetic Epidemiology, University of Utah Research Park, Salt Lake City, UT 84108
(801) 581-5070

The focus of this project is the construction of high-density genetic maps of chromosomes 17 and X and the correlation of these maps with a set of overlapping cloned DNA segments. We have isolated over 70 new restriction fragment length polymorphisms (RFLPs) for chromosome 17 and over 75 for X. The set of available chromosome-17 probes is sufficient to construct a genetic map with an average density of 1 to 2 cM and utilizes the CEPH (Centre d'Étude du Polymorphisme Humain) set of reference linkage families. The set of X markers will permit the construction of a 2- to 4-cM map. Physical mapping of the chromosomes utilizes both naturally occurring translocation break points and a series of selectively isolated "push-pull" hybrids that provide a potentially unlimited series of physical break points from proximal Xp to distal Xq. Physical localization of probes is also facilitated on the X chromosome by studies of males with a variety of disease-associated small deletions, and on chromosome 17 by the existence of deletions associated with partial loss of 17p in some tumor tissues and in the Miller-Dieker syndrome. With the combined application of the above genetic and physical mapping methods, an initial ordering of clusters of DNA probes along each chromosome will be established. The techniques of pulsed-field gel fragment mapping and the isolation of overlapping clones in yeast artificial chromosomes will then be applied to establish an ordered map of all probes and fragments.

Abstracts: Physical Mapping

Human Genome Center, Lawrence Berkeley Laboratory

Charles R. Cantor, C. Bustamante, M. Esposito, J. Gingrich, S. Levene, M. Maestre, R. Mortimer, M. Salmeron, C. L. Smith, and M. Stoneking
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
(415) 486-6800, (FTS) 451-6800

Researchers at the Human Genome Center at Lawrence Berkeley Laboratory (LBL) are developing methodologies needed to complete a physical map and an ordered library of the human genome. A top-down approach will be used by developing yeast artificial chromosome libraries prepared both from total genomic DNA and from specific physically isolated human chromosomes. The immediate goal is to integrate the physical map, as it is developed, with the genetic map by defining the sites on the physical map of cloned and genetically localized genes of specific significance to the Office of Health and Environmental Research (OHER) mission. Several methods for constructing the ordered map will be investigated, some of which include using junction fragments, determining fragment overlap by restriction maps, and employing recombination among artificial chromosomes.

Brief abstracts of the individual projects are listed below.

Optimization of Yeast Artificial Chromosome (YAC) Mapping (M. Esposito, J. Gingrich, R. Mortimer, and C. L. Smith) — The use of larger DNA fragments means that fewer fragments need to be ordered into a map; consequently, the initial major focus has been to obtain clones containing large fragments of DNA from chromosome 21. The most promising avenue appears to be the use of YAC vectors. Different strategies for producing YAC clones are currently being evaluated and optimized. Since the starting material being used for these clones is a hybrid cell line consisting of human chromosome 21 in a background of mouse chromosomes, a major part of the effort is devoted to methods for identifying YAC clones that contain human DNA. These clones are expected to comprise only 1–2% of the total YAC clones from this cell line. The polymerase chain reaction is being evaluated as a new approach to this problem. This identification strategy is based on the expectation that only those clones containing human DNA would be amplified from a known human DNA sequence primer. A method to link up YACs with overlapping sequence information by using recombination is under development, as are new methods for improved DNA transfection into yeast.

New Mapping Methods

Sequence Matching (C. L. Smith and M. Stoneking) — To construct a map, a means of uniquely identifying each DNA fragment is necessary; the strategy at LBL is to use DNA sequences from the ends of the fragments. Knowing a small sequence (50–100 bp) from each end of a large DNA fragment will permit each unique fragment to be identified. Furthermore, matching these DNA sequences with similarly sized DNA

sequences from linking clones (clones that contain the DNA from the ends of two adjacent large fragments) will facilitate map construction. The advantages of this protocol over existing ones is the speed of generating results, the precision of the ordering, the simplicity of data analysis, and the fact that the mapping process generates sequence data as well.

In addition to the traditional means of accomplishing the above tasks, the researchers are also investigating ways of using amplification via the polymerase chain reaction (PCR) to obtain DNA sequences from the ends of large fragments and from linking clones. Ultimately, there are plans to generate linking clones directly via PCR and thereby avoid some of the pitfalls of traditional cloning methods that make completing a map difficult. The PCR-based sequencing strategies are also attractive because they can be readily automated or adapted to existing automated technologies such as DNA sequencers.

Direct Visualization of Chromosomes and DNA Fragments (S. Levene, M. Maestre, M. Salmeron, and C. Bustamante) — Two other second-generation mapping protocols are being investigated. First, hybridization of genes directly on chromosomes that are visualized with confocal microscopy is used to develop physical maps of intermediate resolution. Second, further scanning tunneling microscope (STM) development would produce STM and DNA handling techniques that would allow the nucleotide sequence to be read directly from an isolated fragment of DNA.

Abstracts: Physical Mapping

Physical Maps of Human Chromosomes: Methods Development and Applications

Anthony V. Carrano, Elbert W. Branscomb, Pieter J. de Jong, Emilio Garcia, Harvey W. Mohrenweiser, and Thomas Slezak
Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
(415) 422-5698, FTS 532-5698

The initial goal of this project is to create physical maps of human chromosomes and to correlate them with the genetic map. The physical maps will consist of overlapping cloned DNA fragments (contigs) contained in phage, cosmid, and yeast vectors, all of which span the chromosomes. The project is multidisciplinary, and its components are synergistic. In the past two years, progress has been made in several areas. We constructed new or modified existing vectors to (1) facilitate cloning small amounts of DNA in cosmids, (2) clone *Not* I linking probes in lambda and plasmids, and (3) clone large fragments of DNA as yeast artificial chromosomes (YACs). Several of the cosmid vectors have been transferred to industry. The cosmid vectors were used to construct chromosome-19-specific libraries from flow-sorted chromosomes and from a monochromosomal hybrid. About 10,000 cosmids (about sixfold redundancy) have been arrayed in microtiter trays to form a reference library for chromosome 19. We used the new plasmid and lambda vectors to create a *Not* I linking library of chromosome 19 and have initially isolated about 30 clones. We are currently expanding and characterizing libraries of chromosome 19 in YAC and half-YAC vectors. To construct a set of cosmid contigs for chromosome 19, we developed an automated fluorescence-based strategy for fingerprinting each clone. For this procedure, a robotic system is used to attach fluorophores to the ends of restriction fragments from each cosmid clone. Fragment lengths are determined by using a commercially available laser scanning device to acquire electrophoretic mobility data in real time. Up to four different fluorophores (i.e., four clones) can be run in each gel lane. In the present configuration, this permits us to analyze up to 48 cosmids per gel run. We developed software to process the acquired fluorophore signals, convert the signal data to restriction fragment lengths for each cosmid, and use the fragment length data to compute a statistical measurement of overlap between cosmids. Several thousand cosmids have been processed to date. We have established 6 cosmid contigs that span approximately 600 kb of chromosome 14 and have over 200 contigs for chromosome 19. Five of the chromosome-19 contigs represent known gene loci, and the others are located throughout the chromosome. Contigs are validated by restriction fragment digests and/or by in situ hybridization to metaphase chromosomes. By using large-fragment analysis from pulsed-field gels to close the region of chromosome 19 containing three DNA repair genes and the myotonic dystrophy locus, we discovered that two of the DNA repair genes lie within 260 kb of each other. Finally, we devised a technique, based upon PCR amplification of DNA, to isolate region-specific probes located between human *Alu* repetitive sequences. These probes are being used to identify those

cosmids, from our chromosome-19 library, that span a specific region of the chromosome. As soon as we have processed about 8000 cosmids from chromosome 19 and have initiated the process of contig closure, we will begin to construct cosmid contigs from another human chromosome, probably chromosome 3. The physical mapping effort on this chromosome will be done in collaboration with several other research groups. As the physical maps near completion, we will develop and exploit new sequencing methods to study the molecular architecture of the chromosome and new screening methods that will rapidly evaluate somatic variation and induced genetic change in human populations.

Mapping and Ordered Cloning of the Human X Chromosome

C. Thomas Caskey, David L. Nelson, and David H. Ledbetter
Department of Molecular Genetics, Baylor College of Medicine, Houston, TX 77030
(713) 798-4773

The ultimate goal of this project is the isolation of a complete set of overlapping DNA clones comprising the human X chromosome. This goal will be achieved through several means. A high-resolution pulsed-field gel map of regions of the chromosome will be developed; the regions will begin in Xq28 and extend toward the centromere, in order to assist in placement of clones as they become available. An extensive panel of somatic cell hybrids will be developed to assist in probe isolation and assignment. Rapid isolation of novel X-region-specific fragments will be achieved through the development of a method based on the polymerase chain reaction human-specific *Alu* primers for specific amplification of human sequences from somatic cell hybrid and cloned DNAs will be utilized. Yeast artificial chromosomes retaining large X-region-specific fragments will be utilized for regional isolation and overlap. Finally, a computer database management system will be designed to assist in data handling for these tasks. Initial efforts will focus on the Xq24-qter region with specific emphasis on Xq28, a region with a large number of genetic disease loci.

Physical Map and Overlapping Cosmid Set for Human Chromosome 11

Glen A. Evans, Kathy A. Lewis, Gary Hermanson, Kathryn C. Evans, Jun Zhao, Kimball O. Pomeroy, Carlisle P. Landel, David McElligott, Mary Saleh, James Eubanks, Daniel Kaufman, Ken D. Pischel, Shizhong Chen, Joseph Trotter, Reece Hart, and Grai Andreason
Molecular Genetics Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037
(619) 453-4100, ext. 279

The mission of The Salk Institute is to apply concepts and techniques of modern biology to the solution of human medical problems. Human genome research at The Salk Institute was initiated in 1988 with the support of the Department of Energy (DOE), under the direction of Dr. Glen A. Evans. The recently established Center for Human Genome Research, a closely integrated research group at The Salk Institute, is working in collaboration with several DOE national laboratories within a highly focused program to produce a continuous physical map and overlapping cosmid set for human chromosome 11. Inherent in this approach is strong involvement in the development of novel techniques and strategies for genome analysis.

In the past two years, researchers at The Salk Institute's Center for Human Genome Research have made considerable progress in the development of new cloning methodologies and techniques for genomic analysis and in using these approaches to construct a physical map of human chromosome 11. Extending over about 148 mb, chromosome 11 represents about 4.2% of the human genome. We have constructed, as a pilot project, an initial small set of cosmid clones in a specialized cosmid vector, sCos-1, that allows the rapid determination of overlapping sequences in the collection through the synthesis of directional RNA probes. Over 1000 cosmids mapping in the region from 11q12 to 11qter have been isolated from a cosmid library constructed from a somatic cell hybrid containing a portion of human chromosome 11 in a mouse MEL cell background. These cosmids have been organized in a 36-×-36 array on a nitrocellulose filter; using a novel strategy of overlap determination—referred to as multiplex mapping—that uses pools of clones, we detected 1099 pairs of linked clones in the collection. These pairs have been assembled into 315 predicted cosmid contigs, which are now undergoing analysis by restriction mapping.

Using a laboratory robot, we also devised techniques for automated preparation of cosmid DNA and for restriction analysis. Each of the clones for four rare restriction enzymes—*Not* I, *Bss*H II, *Sfi* I, and *Sac* II—have been mapped, and over 150 *Not*-I-containing linking clones and 37 putative *Hpa*-II-tiny-fragment (HTF) islands have been identified. This automated system is now being used to complete the restriction mapping of the entire collection of cosmids.

Abstracts: Physical Mapping

In collaboration with D. Ward and P. Lichter (Yale University Medical School) and D. Housman and K. Call (Massachusetts Institute of Technology), we have used high-resolution in situ hybridization of single cosmid clones to map selected contigs and landmark clones to chromosomal locations. Produced in the pilot study, the resulting collection of clones, containing over 60% of the 11q12 to 11qter region, is contained in a reference collection now undergoing analysis. To expand this pilot study to a larger collection of clones representing a tenfold redundancy of the entire chromosome 11, we have used a fluorescence-activated cell sorter to purify human chromosome 11 from a somatic cell hybrid, J1, containing a single chromosome 11 in a CHO-K1 cell background, and we are preparing a chromosome-11-specific cosmid library in sCos-1.

This work represents the beginning of a large-scale mapping project to obtain, reference, archive, and link cosmids spanning the entirety of human chromosome 11, a project which may be complemented by studies using pulsed-field gel electrophoresis and yeast artificial chromosomes. During the next year, the Center for Human Genome Research will expand its program to include additional Salk Institute investigators interested in gene recombination and amplification, YAC vectors and mapping strategies, novel in vivo mapping techniques, and identification of recessive oncogenes. The goals are to complete the chromosome-11 physical map and proceed with characterization of genes important to human biology. New methodologies to be utilized in the longer term include DNA sequencing, functional expression in mammalian cells in culture, creation of transgenic mouse strains as models for human disease states, and the identification of functional genes by genetic complementation.

The immediate goals are to (1) improve engineering, robotic, and computational systems for preparation and management of arrayed cosmids and their subsequent processing; (2) expand the arrayed library of chromosome-11 cosmids to 10,000–35,000 members; (3) establish a corresponding repository and database for the distribution of these resources and the correlation of results from other laboratories; (4) pursue detection and characterization of expressed genes, especially those relevant to disease, while completing the physical map; (5) continue the conceptual and practical development of the multiplex strategy including extension of the probe-pooling strategy from two- to higher-dimensional arrays of the cosmids, establishment of the optimal library size, further suppression of effects of troublesome high-copy-number sequences, and exploration of the utility of PCR techniques for probe preparation; (6) develop a correlative approach to integration of chromosome-11 map data acquired through multiplex walking, PFG and linking clone analyses, linkage data obtained through the use of a variable number of tandem repeats (VNTR), RFLP, and minisatellite probes and radiation hybrids; and (7) continue the characterization of DNA/genes adjacent to the chromosome-11 translocation breakpoints obtained from clinical sources and pursue the identification of disease genes mapped to chromosome 11.

Novel Methods for Physical Mapping of the Human Genome Applied to the Long Arm of Chromosome 5

Michael McClelland, Carol A. Westbrook,* Mike Weil, John Hanish, Mike Nelson, Yogesh Patel, Settara C. Chandrasekharappa,* Michelle M. Le Beau,* and Michelle Rebelsky*
California Institute of Biological Research, La Jolla, CA 92037
(619) 535-5476
*Department of Medicine, University of Chicago, Chicago, IL 60637

The goal of this project is to develop and assess new approaches to megabase mapping of a subchromosomal region, specifically applied to chromosome 5, bands q23–31. This region has been selected because, at 25 Mb, it is of manageable size and represents an approach to the larger chromosomes. Our megabase map will consist of restriction sites for enzymes that cleave infrequently and will link a series of probes that have been mapped to 5q. The region is delineated, and the probes sublocalized, by means of hybrids containing translocations and deletions in chromosome 5, some of which have been prepared from leukemic cells of patients that carry chromosome-5 abnormalities. The technology we plan to develop includes (1) enzymologic strategies and (2) methods for the directed production of unique-sequence probes from the region of interest and will include linking clones. The multimegabase strategies have been quite successful: we have developed a reliable method for producing a partial digest of DNA in agarose. In addition, several methylase/*Dpn* I combinations are being evaluated, including one that cleaves a 12-bp specificity. These approaches should generate fragments of over 500,000 bp in the human genome and facilitate the linking of probes. The map will have interesting biological uses because the region contains the gene(s) for radiation/mutagen-induced leukemia, as well as for a variety of growth factors and receptors.

Center for Human Genome Studies, Los Alamos National Laboratory

Robert K. Moyzis, C. E. Hildebrand, R. L. Stallings, and N. A. Doggett
Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-3912, FTS 843-3912

The Los Alamos Center for Human Genome Studies will provide coordination, technical oversight, and direction for the following interdisciplinary elements of the Human Genome Program at Los Alamos: physical mapping, new technology development, and informatics. The center will also develop collaborative research and development programs with the private sector and with other centers for human genome research.

The goals of this project are (1) to develop concepts and to advance technology for genomic physical mapping and (2) to construct a physical map of human chromosome 16 that will include an ordered set of overlapping DNA fragments encompassing the chromosome. The physical map will integrate phage, cosmid, and yeast artificial chromosome (YAC) contigs ordered by repetitive sequence "fingerprinting" with the genetic linkage map, identified gene sequences, and the cytogenetic map into a tool that will allow rapid access to any region of chromosome 16 for analysis and eventual large-scale sequencing. The significance of this work lies in the immediate application of the knowledge and tools (1) to understand human genetic disease, (2) to clarify the molecular bases for genetic disease susceptibility, especially in regard to energy-related chemical or radiation exposures, and (3) to reveal the molecular details underlying long-range chromosome architecture and dynamics.

Genome Organization and Function

Robert K. Moyzis, Julie Meyne, and Robert Ratliff
Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-3912, FTS 843-3912

The ultimate objective of this program is to determine the molecular mechanisms by which higher organisms organize and express their genetic information. Applications of these basic investigations will include the development of novel approaches for (a) detecting of human genetic diseases and (b) measuring the effects of low-level ionizing radiation and/or carcinogen exposure. A combination of biochemical, biophysical, and recombinant DNA techniques is being used to identify, isolate, and determine the roles of DNA sequences involved in long-range genomic order. Currently, major efforts are focused on determining the organization and function of human repetitive DNA sequences. Major findings in the last year included (a) the use of synthetic repetitive DNA oligomers to target in situ hybridization to specific human chromosomes and (b) the isolation of the human telomere. Future studies will be directed toward (a) the further definition and isolation of "functional" repetitive DNA regions and (b) the cloning, in yeast artificial chromosome vectors, of human telomere adjacent DNA fragments. Defining the mechanisms responsible for organizing the mammalian genome, as well as determining the genetic and nonmutational alterations accompanying abnormal phenotypic change, is important to identifying the effects of environmental contaminants from energy-related technologies. Determining the genetic variability in these mechanisms provides a rational basis for establishing thresholds for toxic substance exposures, for making valid cross-species extrapolations, and, ultimately, for identifying individuals at risk.

Developing a Physical Map of Human Chromosome 22 Using PACE Electrophoresis and Large Fragment Cloning

Melvin I. Simon, Bruce Birren, and Hiroaki Shizuya
Biology Division, California Institute of Technology, Pasadena, CA 91106-4107
(818) 356-3944

The goal of this project is to derive a set of overlapping clones covering human chromosome 22. Much of the work involves the development of new or improved methods for cloning large DNA fragments, and for handling, analyzing, and overlapping these clones. To create an overlapping clone map of human chromosome 22, a set of new bacterial artificial chromosome (BAC) vectors will be developed that should support the cloning of large DNA fragments about 200 kb in length. These BAC vectors are based on the *E. coli* F-factor and will be constructed to contain promoters for walking, a multiple cloning site, a cos site for cleavage reactions, rare-cutting sites surrounding the insert, and two selectable markers. A library of chromosome 22 will be constructed in these vectors, in modified YAC vectors, and in cosmids. Source DNA for the libraries will come from hamster/human hybrid cells that contain either intact or deleted chromosome 22. For the hybrids with deletions, the pulse alternating current electrophoresis (PACE) system will be used to separate the deleted chromosome. Human clones will be selected from the libraries by screening with total human DNA. Fingerprints of the clones from the different vector systems will be achieved by partially digesting the cloned DNA and labeling the cos sites with either radioactive or fluorescent tags. The cos sites will be cut with terminase and labeled by a hybridization-ligation reaction. Since each cos end has a different sequence, different oligos can be ligated to each site and a partial-digest map can be created from each end of the clone. The use of fluorescent tags attached by ligation allows the simultaneous use of different fluorochromes at each cos site while separate restriction analysis would be done for radiolabeled oligonucleotides. Detection of the restriction fragments would be performed on the PACE pulsed-field gel electrophoresis system. Based upon the partial digest data, computer algorithms will construct the overlap map.

Techniques for Determining the Physical Structure of Entire Human Chromosomes

Cassandra L. Smith, W. Michels, J. S. Cheng, H. Fang, J. Gingrich, D. Wang, and Y. Wu
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
(415) 486-6800, FTS 451-6800

Large-fragment DNA methods are being used to construct a macro-restriction map of the smallest human chromosome. Isolation of a human telomere yeast artificial chromosome clone enabled the ends of the map to be defined. About 30 single-copy DNA probes with previously assigned genetic map locations along the length of the chromosome are being employed as anchor points. These probes were used to identify corresponding large *Not* I and *Mlu* I DNA fragments by hybridization to pulsed-field gel fractionated DNA restriction digests. The map between the anchor points is being reconstructed by combining several approaches: assigning other bands by using single-copy probes with known regional locations; assigning neighboring bands using the 15 thus-far-isolated chromosome-specific *Not* I linking probes; and interpolating between anchor points using partial digests phased either by Smith-Birnstiel type approaches or by using sites that are polymorphic, when different cell lines are compared, as signatures of particular regions. A series of clones of repeated DNAs are being used to identify all the chromosome-21 restriction fragments present in these hybrid rodent cell lines. These approaches have allowed us to identify about 40 Mb that come from chromosome 21 and to link up *Not* I fragments of at least 8 Mb near the *q* telomere and additional significant regions along the *q* arm. Additionally, these strategies have allowed us to determine that D21S13, the locus most closely linked to the Alzheimer's disease gene on chromosome 21, is in fact located on the same 1.6-Mb fragment as locus D21S16. Although the map is not yet complete, it reveals interesting features such as the uneven distribution of putative genes along the chromosome and a greater-than-expected gradient of enhanced recombination near the *q* telomere.

Correlation of Physical and Genetic Maps of Human Chromosome 16

Grant R. Sutherland, David F. Callen, Valentine J. Hyland, John C. Mulley, and Robert I. Richards
Department of Histopathology, Adelaide Children's Hospital, North Adelaide, South Australia 5006, Australia
011-618-267-7333

The goal of this project is to construct a detailed physical map, which will be correlated with the linkage map, of human chromosome 16. The methods to be used for construction include the following:

(1) A panel of mouse/human hybrid cell lines, which contain only parts of chromosome 16, will be developed. The panel will be achieved by fusing human cells that contain rearrangements of chromosome 16 with mouse A9 cells and selecting for the human APRT gene on the end of the long arm of chromosome 16. The cytogenetic and molecular characterization of the panel will allow detailed physical mapping of cloned DNA sequences and genes that are expressed in the hybrid cells. The cell panel produced should divide this chromosome, which contains approximately 3.3% of the human genome, into about 50 intervals of average-size 2 Mb and thus provide a means of mapping any cloned DNA sequence from chromosome 16 into these relatively small regions. Sequences that map into such a region should then be useful to generate restriction maps using pulsed-field gel electrophoresis. This project should lay the foundation for construction of a restriction map of chromosome 16.

(2) Anonymous cloned DNA fragments of chromosome 16 will be selected from various chromosome-16-specific libraries and mapped using the hybrid cell panel. In selected intervals these fragments will be used to identify restriction fragment length polymorphisms (RFLPs) for which the CEPH (Centre d'Étude du Polymorphisme Humain) panel of families will be typed to correlate the physical and linkage maps. Probes to cloned genes that have been mapped to chromosome 16 will be obtained, and these genes will be physically mapped; where these probes detect RFLPs, their linkage relationships with other cloned segments will be determined using the CEPH families. Probes on chromosome 16 that have been put through the CEPH families and used to generate linkage maps will be obtained from other researchers on a collaborative basis and physically mapped to further define the correlation of the physical and genetic maps of this chromosome.

Automated Methods for Large-Scale Physical Mapping

Tony J. Beugelsdigk and Robert Hollen
Mechanical and Electronic Engineering Division, Los Alamos National Laboratory,
Los Alamos, NM 87545
(505) 667-3169, FTS 843-3169

The preparation of an ordered-clone collection from human chromosome-specific DNA libraries, necessary for both low- and high-resolution physical maps, offers the next challenge in the task of mapping the entire human genome. This research will focus on instrumenting the front-end processes required in constructing a low-resolution physical map, specifically, the ability to propagate automatically and purify human DNA fragments for subsequent analysis and use. We are assembling the necessary automated hardware designed to manipulate, simultaneously, through numerous preparative steps, a large number of samples containing small volumes (0.1–0.5 mL), and, finally, to deliver the samples to solid support filters for binding. The samples will automatically be placed on the solid support in a precisely indexed array for multistage analysis and automated data acquisition.

Based on extensive experience in designing robotic and automated equipment, we anticipate that practical problems in large-scale mapping programs will become evident only after attempts are made to apply new methods to actual map production. For this reason, instrumenting the construction of chromosome-specific physical maps will evolve through a multidisciplinary program. This strategy offers an advantage in that automated devices will become both research and production tools and will be appropriate targets for technology transfer.

Abstracts: Mapping Instrumentation

Human Genome Center, Lawrence Berkeley Laboratory

Charles R. Cantor. C. Bustamante, J. Gingrich, A. Hassenfeld, J. Jaklevic, W. Johnston, J. Katz, W. F. Kolbe, S. Levene, S. Lewis, M. Maestre, and E. Theil
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
(415) 486-6800, (FTS) 451-6800

Researchers at the Human Genome Center at Lawrence Berkeley Laboratory (LBL) are developing innovative techniques in instrumentation and automation to accommodate the size and complexity of the experimental procedures used in physical mapping methods. In addition to improving existing laboratory methods, emphasis will be placed on developing advanced techniques for separating large DNA fragments. Technology for the flow separation of chromosomes will be developed. Modern nuclear radiation detectors or optical and ultraviolet imaging systems will be used to explore methods for direct imaging of electrophoresis gels. The use of differential-polarization imaging to achieve enhanced sensitivity for direct viewing of DNA will be investigated.

Brief abstracts of the individual projects are listed below.

Optimization of Pulsed-Field Gel Electrophoresis (A. Hassenfeld, J. Jaklevic, J. Katz, W. F. Kolbe, and S. Levene) — Engineers at LBL have constructed a test bed for all of the various configurations of PFG electrophoresis. This test bed provides precision control and recording of the conditions within the gel in real time during the run. Optimizing the speed and precision of DNA isolation will, in turn, shorten crucial steps in the mapping process while retaining the high resolution of the current PFG techniques. Among the variables currently being explored are short, intense secondary electric field pulses and combined electric and magnetic fields.

Micromanipulation of DNA (M. Maestre and C. Bustamante) — Some of the information from the PFG studies is being applied to the development of a system for handling and manipulating single DNA molecules. The goal is to develop techniques for isolating specific, single DNA molecules for other procedures such as PCR, cutting by enzymatic or physical means, or direct visualization. The rationale for this technique is to construct a network of electrodes that are about the same size as large DNA molecules (electrodes of 10–20 mm separated by 20–50 mm). A key concept is the use of inhomogeneous fields. If a DNA molecule is to be manipulated in a way that places different sections of the molecule in specific positions in the electrode net, it is essential that these parts experience different forces and different electrostatic fields. The motion of single DNA molecules has been made visible with the use of the intensified epifluorescence microscope after labeling with fluorophores (acridine orange). Direct manipulation of the molecule is then performed through the computer-assisted control of the local electric fields by the microelectrodes. The microelectrodes have been constructed and tested, and DNA molecules can be moved in predictable directions. T4 phage DNA was used in the preliminary testing. The DNA molecules stretched to a length of 49 mm; this length is very close to the length of 52 mm reported for the T4 phase DNA.

Automated Image Analysis (W. Johnston, S. Lewis, J. Jaklevic, and E. Theil) — Most mapping protocols rely upon visualization of DNA. Therefore, development of hardware and software for automatic capture, analysis, indexing, and storage of images is needed to advance the generation of physical maps whether by PFG analysis, confocal microscopy, or STM. The gel imaging system being developed here emphasizes automatic electronic filtering of peaks for band identification. The filter procedure is designed to remove constant or linear background while compensating for the weak signal that may be present from shoulders on better defined peaks. Other aspects of the imaging system include development of a simple-to-use image database and automatic lane compensation. The overall goal is a highly integrated acquisition, analysis, storage, and retrieval system for any image.

Robotics (J. Gingrich, S. Lewis, J. Jaklevic, and E. Theil) — Robotic techniques are being developed to automate and accelerate the labor-intensive steps that currently limit the rate of generating physical maps. This effort is resulting in modification of existing hardware as well as the development of new software. Applications currently being tested for robotic automation include screening yeast colonies for those containing YACs; processing DNA samples for PFG analyses; and collecting and processing DNA samples separated by PFG for further electrophoretic analysis, sequencing, or amplification by PCR.

Genomic Instrumentation

Jack B. Davidson
Instrumentation and Controls Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6010
(615) 574-5599, FTS 624-5599

We are taking a two-level approach to instrumentation needs in DNA mapping and sequencing. On the first level, developments to improve present gel-based techniques are extensions of our approach to filmless autoradiography. Using ultralow-light-level digital television, we detect macromolecules labeled with ^{3}H, ^{14}C, ^{35}S, or ^{32}P directly from dried gels impregnated with a liquid scintillator or by use of an intensifying screen. Originally developed for two-dimensional protein distributions, the method has improved the speed and accuracy of data acquisition and may be useful for imaging the blots found in gene mapping. Upon resolution and field-coverage improvements, large conventional sequencing gels and the blots used in G. M. Church's multiplexing system could be imaged directly. Because light is the detected entity, the basic system can be applied to gels tagged with fluorescent dyes as well as radioactive labels. A related development is a "lensless" radiation microscope for imaging beta particles in in situ hybridization studies and in neuronography. One goal—to visualize radiolabeled genes on chromosomes—requires a 20- to 50-fold improvement in resolution over present capability.

High-Resolution DNA Mapping by Scanning Transmission Electron Microscopy (STEM)

James F. Hainfeld, Martha N. Simon, Stephen G. Will
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
(516) 282-3372, FTS 666-3372

Mapping DNA directly with STEM may complement the fast-growing technology of automatic sequencers. Several tests will be made to determine the feasibility, speed, and reliability of this method. There are several advantages of a direct physical microscope approach to sequencing: (1) very long sequences could be done (i.e., 10^5–10^6 bp in length); (2) if successful, the method could be several orders of magnitude faster than chemical methods; and (3) since long pieces of DNA are used, the problems encountered with repetitive sequences would be circumvented. Preliminary results have been obtained using the following test system. A 622-bp sequence from pBR322 was excised with restriction enzymes and purified. Next, a 128-bp T7 piece was inserted at position 276 (giving a total of 720 bp). The denaturing and renaturing of equal quantities of the 622-bp and 720-bp fragments resulted in 50% formation of heteroduplexes (one 622 strand paired with a 720 strand) and left the extra bases as a single-stranded loop. A 26-mer oligonucleotide that was complementary to a region of the single-stranded insert was synthesized. A chemical modification added a sulfhydryl at the 5′-end of this oligonucleotide, and the undecagold cluster was covalently attached to it. The oligonucleotide and heteroduplexes were then mixed under renaturing conditions and examined using STEM. Control heteroduplexes with no gold clusters show a kink at the position of the 128-bp single-stranded insert, and the total length and length to the insert are consistent with the proposed model. When the gold-oligonucleotide was hybridized, the gold cluster was visible as a tiny bright dot at the "V" vertex of the DNA. The gold cluster was about 10 Å from the base it labels (3 bp), and the accuracy of positioning a base from the end of DNA segments with STEM is 2 bp. A total potential positional accuracy of 3–5 bp should prove useful in the physical mapping of genomes.

Abstracts: Mapping Instrumentation

Thermal Stability Mapping of DNA by Random Fragmentation and Two-Dimensional Denaturing Gradient Electrophoresis

L. S. Lerman, Nashua Gabra, Eric Schmitt, and Ezra Abrams
Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
(617) 253-6658

The thermal stability of the double helix in a standard solvent is fully determined by the base sequence. Within a long DNA molecule, each local region (ranging in length from a few dozen bases up to several hundred base pairs) undergoes a transition from an ordered helix to a disordered, randomized configuration (melting) within a narrow temperature span, typically from 1 to 3°C for the change from 95% helical to 5% helical. It is convenient to characterize the transition in each region, or *domain*, by the Tm, the temperature at which there is a 50-50 equilibrium between the helical and melted forms. Within human genomic DNA there is substantial variation in the local Tm, as much as about 35 degrees, often with distinct and sharp boundaries between adjacent domains. While the pattern and characteristics of this sequence of domains in long DNA molecules is inferred principally by statistical-mechanical theory, the domain content is more directly observable in short DNA molecules by means of absorption spectroscopy as a function of temperature, or by denaturing gradient electrophoresis. Since the Tm of each domain is changed only very slightly by the substitution, addition, or deletion of one or a very few bases, the sequence of domains provides a robust counterpart to the base sequence. The domain map is less sensitive to trivial individual variation (including methylation) than a restriction map and reflects biological function more closely.

In two-dimensional separation of randomly fragmented genomic DNA, each random fragment is identified by X, Y coordinates representing its length and the Tm of the domain with the lowest Tm in the fragment. All fragments in which that domain is the lowest will find a similar gradient level, regardless of their length. This distribution and the response to specific sequence probes provide a means, in principle, for determining the spacing, order, and Tm among those domains that have a lower Tm than the average and those with the highest Tm. It will provide measurements of the nucleotide distances between each of these domains and any arbitrary set of sequence identifiers or probes.

Our current effort is concerned with refining and calibrating various aspects of the two-dimensional denaturing gradient technique and related procedures. These efforts include (1) using iron-peroxide nicking and S1 cleavage for the preparation of fully random distribution of fragments from lambda DNA and yeast artificial chromosomes containing long human genomic inserts; (2) reducing the breadth of bands produced by very long DNA molecules in the denaturing gradient; (3) analyzing the band broadening observed when the domain with lowest Tm is surrounded by higher melting domains; and (4) developing optical, mathematical, and computing procedures for calibrating gel photographs or autoradiographs in terms of quantitative, point-to-point distributions of DNA.

New Approaches to DNA Mapping: Synthetic Endonucleases

Betsy M. Sutherland and Gary A. Epling
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
(516) 282-3293, FTS 666-3293

Recognition and mapping of functionally important DNA regions (e.g., regulatory and coding regions and initiation sequences) can be greatly facilitated by specific DNA cleavage at such sites. Synthetic endonucleases, able to cleave at regions of functional importance, will be created by coupling DNA site-specific binding proteins via linker arms to light-activatable cleaving moieties; specific binding function is provided by the DNA binding protein; cleavage activity is provided by the activatable cleaving molecules. A prototype system of Rose-Bengal (RB) coupled via a hexanoic acid linker to a DNA lesion site-specific monoclonal antibody will be developed for other specific DNA-binding proteins, including the T7 RNA polymerase and mammalian transcription initiation factors.

Coupling of RB-hexanoic acid to T7 RNA polymerase via 1-ethyl-3-(3-dimethylaminopropyl) coupling (EDC) was found to yield RB-tagged polymerase, which can specifically bind to a plasmid containing a T7 promoter. However, the conditions for EDC were sufficiently stringent to result in low final yields of active polymerase. We designed and synthesized an RB triethylene glycol succinate activated ester, which can be added to polymerase under buffer conditions optimal for enzyme stability. Levels of the RB addition that yielded maximum specific binding of the polymerase to T7 promoter sites were determined. Preliminary results indicate that the RB-labeled T7 RNA polymerase can mediate the cleavage of T7 DNA to a level of at least 2.4 cleavages per DNA molecule. The sites of cleavage by the tagged polymerase are being determined.

Quantitation in Electrophoresis Based on Lasers

E. S. Yeung
Environmental Sciences Program, Ames Laboratory, Ames, IA 50011
(515) 294-8062

The goal of this project is to develop a novel laser-based imaging technique for quantitation in gel electrophoresis and in capillary electrophoresis. No stains are required; thus cost-efficiency, reliability, convenience, and speed of processing are increased. The fact that the scanning technique uses no mechanical parts adds to the positional accuracy (resolution) of the measurement. The research is based on indirect fluorometry and acousto-optic imaging. In indirect fluorometry, a fluorescing ion is used to elute the sample and thus produce a large fluorescence background signal throughout the gel. When one of the components of the samples appears, the fluorescing ion is displaced; a lower fluorescence signal will then be observed. Since electrophoresis is based on charged species, electroneutrality requires a one-to-one displacement of the fluorescing ion. This negative signal allows nonfluorescing species to be detected with the high sensitivity normally associated with fluorescing species only, and without staining. The response should be uniform and predictable because it is derived from the same fluorescing ion. Preliminary results indicate that indirect fluorometry is feasible for monitoring nucleotides and DNA fragments. Applications to DNA mapping and sequencing will be pursued.

Scanning Tunneling Microscopy of DNA

Rodney L. Balhorn and **Wigbert Siekhaus**
Biomedical Sciences Division, Lawrence Livermore National Laboratory,
Livermore, CA 94550
(415) 422-6284, FTS 532-6284

Researchers at Lawrence Livermore National Laboratory, and at other institutions, have recently shown that scanning tunneling microscopy (STM) can be performed on the DNA molecule with angstrom resolution. In addition, STM in the spectroscopic mode (scanning tunneling spectroscopy or STS) has been used to characterize the electronic structure of semiconductor substrates and their interaction with molecules adsorbed on such substrates. The goals of this project are (1) to develop the instrumentation and techniques required for imaging naked double- and single-stranded DNA at or near atomic resolution using STM and (2) to devise methods for obtaining spectroscopic information with STM that allow us to distinguish between the four bases and to sequence DNA. To accomplish these goals, the four nucleotides and various single- and double-stranded DNAs will be imaged, after electrophoresis, onto graphite and other substrates. Experiments will be performed to determine how specific counterions and the hydration state of the deposited DNA affect imaging. To identify individual bases and specific molecular tags, measurements of synthetic and tagged sequences of known length will be made so that various types of spectroscopy (work function, laser-enhanced vibration, and photon emission) can be performed on the molecules. Our application of STM and STS to the analysis and sequencing of DNA should directly impact the progress of the human genome effort by eventually providing a new, electronic method for sequencing DNA at three orders of magnitude faster than existing methods. The techniques and instrumentation developed during the course of this project will also be directly applicable to the analysis of biological samples in general.

Abstracts: Sequencing Technologies

Transposon-Facilitated DNA Sequencing

Douglas E. Berg, Clara M. Berg, and Henry Huang
Department of Microbiology and Immunology, Washington University,
St. Louis, MO 63110
(314) 362-2772

Two types of derivatives of transposon Tn5 will be constructed to facilitate the sequencing of cloned DNAs. One type is designed for in vivo insertion at many sites in DNAs cloned in lambda phage and in cosmids and other plasmid vectors, and for sequencing in both directions from each site of insertion. The other type will be embedded in cosmids and used to generate nested deletions with one variable end point in the cloned DNA and one end point fixed at a transposon end; the set of nested deletions will similarly permit sequencing of the entire stretch of cloned DNA without need for subcloning of random fragments. The Tn5 element for insertion into lambda and cosmids will contain a supF (suppressor tRNA) gene as a selectable marker. Its transposition to lambda will be selected by plaque formation, while its transposition to plasmids will be selected by suppression of a chromosomal amber mutation. Constructs for making nested deletions by intramolecular transposition will contain a Tn5 transposase gene, whose expression is turned on by IPTG, so that deletions can be made at will. The construct will also contain a conditionally lethal gene for selection of the transposition-induced deletions. Deletion-generating Tn5 derivatives will be constructed as complete cosmid vectors for the construction of new recombinant DNA libraries and as cassettes for insertion into cosmids that already contain cloned DNAs. Both types of Tn5 derivatives will be adapted for multiplex sequencing.

Computer-Assisted Multiplex DNA Sequencing

G. M. Church, G. Gryan, S. Kieffer-Higgins, L. Mintz, M. J. Rubenfield, and M. Temple
Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02138-3800
(617) 732-7562

Several laboratories are sequencing genomes (ranging from 1 to 15 Mbp) from each phylogenetic kingdom. The genome closest to completion is *E. coli* (20% of 4.7 Mbp). These sequences will define consensuses for classes of protein domains, evolutionary conservation, and change. While participating in this quest, we have developed a new multiplex DNA sequencing method [Church et al., *Science* **240**, 185–188 (1988)]. In multiplex DNA sequencing, 480 sequencing reaction sets, each tagged with specific oligonucleotides, are run on a single gel in 12 pools of 40 and transferred to a membrane. We hybridize 75 such membranes simultaneously. The resulting sequence film images are digitized, and sequence interpretations are superimposed on the enhanced two-dimensional images for editing. The computer program (REPLICA) uses internal standards from multiplexing to establish lane alignment and lane-specific reaction rules by discriminant analysis. The automatic reading phase takes one hour per film (3 kb) on a Vaxstation. Images with overlapping data can be viewed side by side to facilitate decision making. Hash-table-based routines for linking up shotgun sequences in the megabase range are compatible in speed with the rest of the software.

Abstracts: Sequencing Technologies

Sequencing of Megabase Plus DNA by Hybridization: Method Development

R. Crkvenjakov, R. Drmanac, Z. Strezoska, and I. Labat
Center for Genetic Engineering, Vojvode Stepe 283, P.O. Box 283, 11000 Belgrade, Yugoslavia
38-11-491-391

The DNA sequence of a genome can be constructed by joining shorter overlapping oligomer sequences of DNA. Sequencing by hybridization (SBH) is an implementation in which the oligomer sequences are accessed through a DNA hybridization methodology. Oligomer probes are used to ascertain the presence or absence of complementary sequences in arrayed clones, with the library of overlapping clones representing the genome. The sequence data gathered earlier provide for the ordering of the library. The data on the unique regions of overlapping clone pairs have a particularly important role and serve for the resolution of sequence branch ambiguities that would otherwise arise in the developing sequence of single clones. Through computer modeling, an optimal system configuration is defined that specifies the family of 100,000 oligomer probes needed and the characteristics of the library. For practical application, probe-template hybrids that are perfectly base-paired must be distinguishable from those with a base-pair mismatch. Now that an understanding of the requisite thermodynamics has been achieved and the resolving capacity demonstrated, oligomers—at least as small as octamers—can serve as probes. These size reductions are important because costs of oligomer synthesis are thereby greatly decreased. The continuing development program includes optimization of probe labeling, library cloning, and management procedures; measurement of the actual error rate in data acquisition; and a stringent test of the theory. The latter is a computer simulation of data processing and sequence assembly for a DNA of yeast genome size. A scheme for extensive miniaturization of the probe–clone array system is under development. The aim of the project is to provide experience for decision making on whether or not to proceed to a sequencing pilot plant stage.

*This project is being supported under terms of a Scientific and Technological Cooperation Treaty between the United States and Yugoslavia, with Yugoslavia providing the majority of the funds.

Rapid Preparation of DNA for Automated Sequencing

John J. Dunn and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
(516) 282-3012, FTS 666-3012

A strategy that uses a library of oligonucleotide primers of length eight, nine, or ten has been developed for direct sequencing of cosmid DNAs. The statistics of priming indicate that a library sufficient for determining the sequence of hundreds of thousands of different cosmids could be readily assembled. This strategy would greatly reduce the cost and effort of human genome sequencing. Any needed primer would be instantly available at a cost of considerably less than 0.1 cent per nucleotide of sequence obtained. Mapping, subcloning, or preparation of multiple DNA samples would not be necessary, and the wasteful redundancies of random sequencing strategies would be eliminated. The success of this strategy requires only that a considerable fraction of all octamers, nonamers, or decamers be able to prime selectively. Work is under way to establish conditions where this will be the case. The transposon gamma-delta is being modified to carry genetic signals that enable bacteriophage T7 to replicate and package plasmid DNAs. The ability of such an element to insert these signals into a cosmid DNA and thereby to facilitate preparation of the DNA for sequencing is being tested using a cosmid that carries a complete genomic copy of the receptor gene for polio virus.

Abstracts: Sequencing Technologies

Scanning Tunneling Microscopy

T. L. Ferrell, R. J. Warmack, and Dave Allison
Health and Safety Research Division, Oak Ridge National Laboratory,
Oak Ridge, TN 37831
(615) 574-6214, FTS 624-6214

This project includes the operation and continued development of scanning tunneling microscopes for basic physics research related to health and environmental problems. The primary focus of this research is the use of scanning tunneling microscopes in sequencing the human genome. A scanning tunneling microscope with single-atom resolution is used to image atomic structure on surfaces, to alter atomic positions, and to probe the dynamical phenomena caused by collective electron motion and motion of ions. Development includes extending capabilities to a wider range of materials, to identification of atomic species, and to studies of biological samples. Methodologies are developed for atom-by-atom studies of biological molecules important in understanding the human genome and in providing other applications in surface science.

Multiplex DNA Sequencing

Robert Weiss and **Raymond F. Gesteland**
Howard Hughes Medical Institute, Department of Human Genetics, University of Utah Medical School, Salt Lake City, UT 84112
(801) 581-5190

We have developed a method for rapid DNA sequencing that employs multiplex probing, as originally described by Church and Gilbert. A large number (25–50) of DNA samples, each cloned in an equal number of special vectors, are prepared together and as a mixture are sequenced using dideoxy sequencing from a universal primer whose sequence is carried by each vector. After conventional separation by gel electrophoresis, the mixed DNA pattern is transferred to a membrane. Each of the individual sequences is then revealed by repetitive rounds of probing, washing, reading, stripping, and reprobing with labeled oligonucleotide, each of which is unique for a sequence in one of the vectors. By fixing a number of such membranes in a drum, the ^{32}P-labeled probing can be done (without handling the membranes) to obtain 10,000–20,000 bases of sequence in each cycle with a cycle time of 1–2 days, including time for autoradiography—with minimal labor. By using a set of transposons containing appropriate primer and identifier sequences, we are developing, with help from Diane Dunn, a method for creating in vivo random subclones that are ready for multiplex sequencing. An optical lab has been set up by Jeff Ives and Achim Karger with the help of Joel Harris (Chemistry Department) to compare the feasibility of fluorescence and chemiluminescent tags for the multiple probes. The efficiency of multiplex sequencing has been diminished by a bottleneck at the step of reading autoradiograms. To solve this problem and to deal with data that might come from charge-coupled display (CCD) images of the fluorescent or chemiluminescent membranes, a research group—including Jeff Ives, Mike Murdock, and Tom Stockham and Neil Cotter (Electrical Engineering Department)—has been assembled. Harold Swerdlow is investigating the feasibility of using gel electrophoresis in microbore capillaries (70 mm) to do sequencing.

Abstracts: Sequencing Technologies

DNA Sequencing Using Stable Isotopes

K. B. Jacobson, H. F. Arlinghaus,* G. M. Brown, R. S. Foote, F. A. Larimer, R. A. Sachleben, N. Thonnard,* and R. P. Woychik
Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077
(615) 574-1204, FTS 624-1204
*Atom Sciences, Inc.

This project utilizes a new DNA-sequencing approach that could increase the rate of sequence determination 100-fold or more as compared to current methods that use radioisotopes. In this procedure, stable isotopes of sulfur, tin, iron, mercury, and other elements can be used to label DNA itself or oligonucleotide probes that will be used to locate DNA fragments after electrophoresis. Resonance ionization spectroscopy (RIS) will be used for localization and quantitation of these elements and all their stable isotopes in an optimal fashion. The sensitivity of RIS, as implemented in Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS), should be comparable to that of radioisotopes and fluorescent labels that are in current use. Furthermore, the multiplex method of DNA sequencing will be adapted for use with stable isotopes so that 40 labels can be used simultaneously. This adaptation will require that the maximum number of stable isotopes of a given element be attached to a series of oligonucleotides in a stable configuration, so that the label does not interfere with accurate hybridization. Several chemical approaches are presented that should lead to properly labeled probes that meet these criteria. Current state-of-the-art SIRIS will be used to demonstrate this new sequencing approach, and the design of the instrument will be modified to adapt it specifically to DNA sequence determination. The use of radioisotopes is eliminated along with the attendant problems related to radiation exposure of personnel, the prohibitive costs of radioactive waste disposal, and the need to cope with the short half-lives of reagents that contain radioisotopes. A patent application is pending for this technique. To carry out the goals of this project will require close collaboration of physicists, chemists, and molecular biologists.

Sequencing of Linear Molecules

Joseph M. Jaklevic and W. F. Kolbe
Instrumentation Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
(415) 486-5647, FTS 451-5647

Innovative methods for determining the linear sequence of nucleic acid bases along mapped fragments of human chromosomal DNA are required for the efficient implementation of human genome sequencing. This project involves the investigation of physical analytical methods for determining the base sequence of DNA. The initial approach focuses on the development of techniques for manipulating individual DNA molecules or ordered arrays of identical molecules in a manner that will allow sequencing of the individual bases using direct spectroscopic methods. An intermediate step will be an attempt to combine the manipulation of ordered arrays with solid-phase restriction enzyme chemistry to achieve an advanced method for physical mapping of intermediate-sized fragments. Since the spatial resolution and analytical sensitivity required for DNA sequencing are significantly beyond existing capabilities, the feasibility of combining several innovative technologies will be explored. Initially, we will investigate low-temperature solid matrices as media for both physical manipulation and chemical isolation of individual molecules. Subsequent alignment of the DNA strands using electromagnetic fields will be investigated using direct imaging methods such as scanning tunneling microscopy and fluorescence microscopy of labeled molecules. Methods for attaching cloned DNA fragments to appropriate substrates will also be incorporated into the matrix studies. Spectroscopic methods applicable to the detection of the DNA molecules at the required level of sensitivity will also be explored.

Abstracts: Sequencing Technologies

Advanced Concepts for Base Sequencing in DNA

J. H. Jett, R. A. Keller, J. C. Martin, E. B. Shera
Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-3843, FTS 843-3843

We are addressing the problem of rapidly sequencing the bases in large fragments of DNA. The ideas presented represent the combined effort of a multidisciplinary task force composed of physicists, physical chemists, cellular and molecular biologists, and organic chemists. To reduce mapping requirements, the emphasis is on sequencing methods that are rapid, require little DNA, and are capable of sequencing large fragments. After evaluation of several physical approaches to sequencing, the decision was made to proceed with a modified-flow-cytometer approach that employs laser-induced fluorescence to detect individual fluorescent molecules. A large fragment of DNA, approximately 40 kb in length, will be labeled with base-identifying tags and suspended in the flow stream of a flow cytometer capable of single molecule detection. The tagged bases will be sequentially cleaved from the single fragment and identified as the liberated tag passes through the laser beam. We are projecting a sequencing rate of 100 to 1000 b/s on DNA strands approximately 40 kb in length.

Ultrasensitive Fluorescence Detection of DNA in Sequencing Gels

Richard A. Mathies and Konan Peck
Chemistry Department, University of California, Berkeley, CA 94720
(415) 642-4192

Our goal is to develop an ultrasensitive apparatus for the detection of fluorescently labeled DNA fragments and probes in sequencing gels. A detection system has been developed that is sensitive enough to observe the fluorescence emitted by individual fluorescent molecules.[1] The idea is to detect the bursts of fluorescence arising from the passage of individual molecules through a tightly focused laser beam. A theory for the signal-to-noise ratio in laser-induced fluorescence has been developed and used to select the optimal laser power and molecular transit time (Mathies et al., in preparation). This theory has been satisfactorily tested by studying the fluorescence from phycoerythrin as a function of the incident laser power and the transmit time of the molecule through the laser beam. Using these optimized conditions, we have observed the fluorescence of individual molecules of phycoerythrin in a flow system as they pass through the laser beam. The calculated autocorrelation function of the detected photons clearly shows that we are observing correlated bursts of fluorescence from individual molecules. Also, the frequency of these bursts is linearly related to the concentration as expected for single-molecule events. Using a hard-wired version of our single-molecule detector, we have been able to detect phycoerythrin at concentrations as low as 1×10^{-15} *M* (Peck et al., *Proc. Natl. Acad. Sci. USA*, in press). This detector is three orders of magnitude more sensitive than conventional fluorescence detection systems. A patent on this on-line, single-molecule counter has been filed.[1] We are now in the process of applying this new technology to the detection of individual molecules on substrates and in gels.

[1]R. A. Mathies, K. Peck and L. Stryer, "High Sensitivity Fluorescent Single Particle and Single Molecule Detection Apparatus and Method," patent filed.

Informatics

Abstracts

The Human Genome Information Resource

Christian Burks, T. Michael Cannon, Michael J. Cinkosky, James W. Fickett, C. Edgar Hildebrand, James H. Jett, Rebecca J. Koskela, Frances A. Martinez, Debra Nelson, Robert M. Pecherer, Karen R. Schenk, Robert D. Sutherland, **David C. Torney**, and Clive C. Whittaker
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-7510, FTS 843-7510

Overview. The goal of the Los Alamos National Laboratory (LANL) Human Genome Information Resource (HGIR)—to develop information management and analysis tools for physical mapping data—reflects an interest in linking databases from two or more biological disciplines with the long-term objective of extending similar research to other related data sets such as nucleotide sequences and genetic maps.

The HGIR project has focused initially on developing strategies and tools for facilitating the flow of data from electrophoretic gels into computers and, once in the computer, into various formats for access and analysis. The reasons for this focus are the desirability of examining and manipulating "real" experimental data and HGIR's close ties with the LANL Life Sciences Division experimental group, who are currently developing a cosmid-based physical map of human chromosome 16. The flow of data begins with the digitization and processing of electrophoretic images. Data corresponding to the clones analyzed on the electrophoretic gels are then passed into a computerized database—the Laboratory Notebook (see below). The availability of these data for subsequent analysis of clone fingerprints leads to the development of contig maps and—eventually—comparison to other, related data sets that will allow for higher-order assemblies and ordering of contigs.

At the end of this flow path, the need for more sophisticated data management, analysis, and interface tools becomes evident; therefore, the thrust of the HGIR project will shift to the development of these tools and their use. To facilitate the cross-linking among multiple levels of physical mapping data, as well as between physical maps and other related data sets (e.g., sequences and genetic maps), our data structure design work is being undertaken with this emphasis (and the future extension to yet unrecognized data structures) in mind. Finally, we plan to design an on-line system and set of interfaces to provide these data and tools to a user community more broad than that of the current in-house activity.

Laboratory Notebook. The Laboratory Notebook database is designed to be useful, with little modification, for most mapping strategies. Representing an electronic (and advantageous) alternative to the ubiquitous paper notebook traditionally used in the laboratory to maintain experimental data, the Laboratory Notebook was designed to manage data resulting not only from experiments, but also from information relating to

the corresponding materials, methods, and procedures. The database resides in Sybase, a relational database management system, on a Sun 4 computer operating under UNIX.

We have developed tools to support the flow of data from the laboratory into the database and, once in the database, into various forms suitable for analysis and presentation. For example, electrophoretic gel data (e.g., DNA fragment sizes) are transferred directly to the database from the BioImage Visage 110, an image-processing workstation. Fingerprint annotation is developed and attached to the DNA fragment data items through a similar mechanism. Reports from these data are created for direct input to fingerprint analysis software.

The user interface was designed to present a conceptual view of the data without burdening the user with the need to understand the structure of the database and details of data storage. Developed using the Sybase's APT-forms, the Laboratory Notebook requires very little training to use.

Abstracts: Informatics

Human Genome Center, Lawrence Berkeley Laboratory

Charles R. Cantor, W. Chang, D. Gusfield, M. Hutchinson, W. Johnston, G. Lawler, P. Li, V. Markowitz, D. Naor, F. Olken, and C. L. Smith
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
(415) 486-6800, (FTS) 451-6800

Researchers at the Human Genome Center at Lawrence Berkeley Laboratory (LBL) are developing computational tools needed to analyze the map, clone, and sequence data generated from human genome research and to provide the technical foundation for the construction of a Human Genome Information System (HGIS). Plans call for the development of logical data models, physical structures, access methods, and query facilities for map and sequence data. Also to be investigated are methods for fragment overlap detection, map assembly, and approximate sequence and pattern matching.

Brief abstracts of the individual projects are listed below.

Chromosome-21 Database (W. Johnston, F. Olken, and C. L. Smith) — The current goal in database management for maps and sequences is a prototype, integrated chromosome-21 database and mapping system. The system is initially targeted at groups involved in developing the physical map of chromosome 21 but can be extended to groups studying other aspects of chromosome 21. This system will provide access to various types of physical, cytogenetic, and genetic maps of chromosome 21; a bibliography of chromosome-21 literature; map construction algorithms; and graphic tools for drawing, displaying, and editing maps.

Computer-Assisted Laboratory Notebook (W. Johnston, M. Hutchinson, V. Markowitz, F. Olken, and C. L. Smith) — Image storage and indexing for generating physical maps requires interfacing a data management system with a mass storage facility. The data management group has made progress in using an extended entity relationship as a basis for a data management system that will combine elements of a graphical and textual interface with data interchange between numerous repositories of data. This system will allow data to be loaded in such a way that the data can be updated, edited, reviewed, queried, and manipulated with ease. The computer-assisted laboratory notebook also provides a means of tracking individual experiments, handling inventories, and integrating data for the development of physical maps.

Theory Group (W. Chang, D. Gusfield, G. Lawler, P. Li, and D. Naor) —Theoretical work is being done on two problems: string matching for sequence alignment and map assembly for probed partial digestion experiments. The map assembly effort has looked at backtracking algorithms, including various special cases (e.g., probes on adjacent fragments). In the coming year these researchers expect to continue their investigations and look especially at incorporating polymorphism data and multiprobe experiments. The string-matching work has yielded a fast algorithm for approximate string matching with a bounded number of errors. Also expected to be improved are the error-handling abilities of the algorithm.

GnomeView: A Color Graphics Interface to the Human Genome

Richard J. Douthart, Victor B. Lortz, and David A. Thurman
Battelle Pacific Northwest Laboratory, Life Sciences Center, Richland, WA 99352
(509) 375-2653

Battelle Pacific Northwest Laboratory is developing an interactive, graphical computer interface to information and data being generated by the human genome project. Color graphic representations of genetic objects and maps provide the user with a sense of topology and reveal patterns that are otherwise difficult to detect. The GnomeView software allows a user to look at the entire human karyotype and manipulate graphical representations of genetic and physical maps and sequences. GnomeView supports the creation and display of maps, which are rendered in graphics on a workstation using the X Window System. The user can zoom, scroll, and select graphical objects. Descriptive information is available by selecting particular objects or groups of objects. Textual references may also be followed using GnomeView's hypertext features. GnomeView includes a network-model database that permits natural and efficient representation of the hierarchies in genomic information. Landmarks, features, and blocks of sequence information are stored as objects in the database. Compound objects and maps are created by reference to other objects. Specific attention has been paid to designing database search algorithms and user interface techniques that scale well to high data volumes. We expect that eventually the GnomeView software will contribute substantially to obtaining closure of the various types of maps currently being generated. For instance, a restriction enzyme map can be updated from analysis of sequence information that sets exactly the position of cut sites. Different maps covering the same region can be compared visually and analytically for inconsistencies.

Abstracts: Informatics

Identification of Genes in Anonymous DNA Sequences

C. A. Fields and C. A. Soderlund
Center for Advanced Computing in Molecular and Cellular Biology, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001
(505) 646-2848

The objective of this project is the development of practical software to automate the identification of genes in anonymous DNA sequences from the human and other higher eukaryotic genomes. A prototype automated sequence analysis system, **gm2**, has been implemented in the programming language of C for Unix version 4.2. This system accepts for input: DNA sequences; consensus matrices for locating splice sites, translational start sites, and polyadenylation sites; match-quality cutoffs for consensus searches; and base frequency and codon usage standards for coding regions and introns. It produces, as output, schematic models of the possible genes contained in the sequence that show the locations of the coding sequences, introns, and control signals. Extensively tested on sequences in the 10-kb size range containing known genes of up to 10 exons, **gm2** is capable of generating complete and correct analyses showing all possible alternative splicing patterns. Run times for such analyses on a Sun 3/60 workstation range from less than 1 min to about 45 min and depend on the stringency of the search parameters used. Current effort is focused on implementing procedures for analyzing sequences that contain only partial genes and on implementing a more efficient algorithm for first-pass analysis using low-stringency parameters.

Special-Purpose VLSI-Based System for the Analysis of Genetic Sequences

T. Hunkapiller, M. Waterman, R. Jones, M. Eggert, E. Chow, J. Peterson, and **L. Hood**
Department of Biology, California Institute of Technology, Pasadena, CA 91125
(818) 356-6408

The current size of genetic sequence databases means that extensive similarity analyses based on robust mathematical models require large, expensive computer hardware not generally available to most investigators. We are currently involved in developing a hardware alternative—a VLSI-based system (i.e., very large scale integrated)—that will give most laboratories access to these rigorous algorithms at the level of affordable workstations and/or PCs. We are designing a board-level biological information signal processor (BISP) coprocessor assembly. BISP is a systolic implementation of dynamic programming methods for the determination of local sequence similarities and alignments. Each BISP board will include an array of BISP processor chips responsible for performing the dynamic programming methods and a tightly coupled coprocessor (i860) that will provide complete alignments to the host central processing unit (CPU) from the scoring information and locations provided by the systolic array. Difference tables and gapping penalties are completely user-definable. Also, there are no arbitrary limits on sequence lengths of either the query or database sequences, and searches can be made simultaneously for multiple query sequences (depending on the size of the sequences and the systolic array). The array will process at up to 20 megacharacters per second. At this maximum speed, a query sequence the size of the systolic array could be compared against all of the current GenBank® (both orientations) in under 3 s. The size of the array is extensible and will range from about 500 processor cells to several thousand. The BISP chip is being implemented in 1 mm cMOS technology, based on custom standard cell methods. Each chip will have multiple identical processor cells, local control circuits, and a results cache. The nearly full-custom processor cell design is nearing completion, and the control circuit logic and floor plan have been generally detailed. Our present schedule calls for tested design vectors to be completed by late fall and handed over to the fabrication service for prototype chip delivery by winter 1990.

Abstracts: Informatics

Human Genome Management Information System

Betty K. Mansfield, Judy M. Wyrick, **John S. Wassom**, Po-Yung Lu, Mary A. Gillespie, and Sandy E. McNeill
Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6050
(615) 576-6669, FTS 626-6669

The Human Genome Management Information System (HGMIS), sponsored by the U.S. Department of Energy (DOE) at the Oak Ridge National Laboratory, has the following roles in the Human Genome Program: (1) to assist the DOE Office of Health and Environmental Research (OHER) in communicating issues relevant to human genome research to DOE contractors and grantees and to the public and (2) to provide a forum for exchange of information among individuals involved in genome research or the development of instrumentation and methodologies to implement genome research. To fulfill these communications goals, HGMIS is producing technical reports, DOE Human Genome Program reports, a quarterly newsletter, and an electronic bulletin board. These documents/facilities are available to all interested persons upon request. The first technical report will assess instrumentation and methodology development relevant to DNA mapping and sequencing. The DOE Human Genome Program reports will contain information on human genome research and development activities supported by the DOE program, as well as background information. The *Human Genome Quarterly* newsletter features technical articles, meeting reports, a calendar of genome events, announcements, and other information relevant to genome research. Accessible via modem through direct dial or via the user's host mainframe computer network, the electronic bulletin board contains information organized by categories (e.g., menu, general information, news, and comments from OHER; summaries and highlights of research projects; meeting announcements/calendar; literature highlights; DOE Human Genome Program contacts; and international activities). HGMIS welcomes comments, suggestions, and contributions from the genome research community.

Applications of Logic Programming and Parallel Computation in Genetic Sequence Analysis

Ross Overbeek, Ian Foster, and Steve Winker
Mathematics and Computer Science Division, Argonne National Laboratory,
Argonne, IL 60439
(312) 972-7856, FTS 972-7856

Several applications of logic programming and parallel computations to genome research problems are being pursued in close collaboration with G. Church (Harvard Medical School), C. Woese and G. Olsen (University of Illinois at Urbana), and M. Liebman (Amoco Technology Co.). (1) For interpretation of sequencing gels, a program has been developed that mimics human expert behavior and includes extensive backtracking to explore sets of alternative interpretations. (2) For multiple sequence alignment, a new algorithm has been devised. It implements several heuristic methods to produce alignments of up to 500 sequences with lengths of 2000 nucleotides. With the use of bilingual programming (logic programming with C routines that handle well-defined computationally intensive subcomputations), prototype algorithms could be rapidly tested and executed on multiprocessors. (3) For prediction of secondary structure manifesting in RNAs, an approach based on covariance analysis has been developed that heavily utilizes multiple sequence alignment. Current applications are the explorations of possible interactions between 5S, 16S, and 23S RNAs. (4) For the problem of constructing databases with very diverse data types, uses of logic programming are being explored. To increase capacities to handle large data sets in disk, usual logic programming environments have been extended by the addition of predicates. The result is a prototyping environment that can manage the increasing volume of sequence data. (5) For comparisons of relative cost and performance, sequence similarity search algorithms are being compared on several commercially produced multiprocessors. These include the shared memory MIMD (multiple instruction, multiple data) machines, SIMD (single instruction, multiple data) machines [similar to the Connection Machine™ (Thinking Machines, Inc.) and DAP™ (Active Memory Technology)], and distributed memory machines. Logic programming technology is also being utilized here to explore solutions that are portable over a wide range of machine environments.

Abstracts: Informatics

Computational Support for the Mapping of Complex Genomes

Karl M. Sirotkin, Daniel S. Caugherty, David C. Torney, and Frederic R. Fairfield
Center for Human Genome Studies, Group T-10, Mail Stop K710, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-7510, FTS 855-0479

At Los Alamos National Laboratory (LANL), we are determining how to construct contigs (contiguous segments of DNA) from fingerprint data and are making tools to evaluate a number of methods to optimize the rate of contig extension and completion of a map. From the fingerprint data generated at LANL, contigs are assembled by using calculated overlap-probability values for pairs of cosmid clones. These overlap probabilities are determined from a statistical model that incorporates the positions of the fingerprint data on a gel and the reproducibility of this data. Extensions of these results will include calculations of the likely structure of a contig (including clone positions and partial restriction maps) and analysis of the position-dependence of the chromosome-16 fingerprint data.

To evaluate mapping methods, we have created a set of modular programs that determine the limitations on contig size when a particular mapping strategy is used. A person using these programs can create a test genomic segment, generate clones from this segment, extract fingerprint data from each clone, test various strategies for determining pairwise overlap between clones, reassemble the genome from these overlaps, display the resulting contigs, and evaluate the success of the reassembly. The programs have been structured to evaluate many mapping strategies and to assemble contigs from current data. To facilitate the use of these programs by others, both program and data files are readable by the user, and the same symbolic parameter names are used throughout the data and program files. In addition, we have designed the program structure so that it is easy to tailor the installation for individual users.

These modeling programs have been implemented in two phases. The first phase used only exact fingerprint data; in the second phase, there were controlled levels of error in the fingerprint data. By using exact data, four different strategies reassembled similar percentages of a genome segment into contigs, although the exact contigs were different. By combining information from all of these methods, the coverage of the genome by contigs increased. With errors in the fragment lengths, it is not yet clear which strategies for reassembly of fragments into contigs make better use of the data. Small changes in the experimental errors seem to have large effects on contig generation. With this project, we hope to optimize map construction rates by making contigs, monitoring our progress, and circumventing potential problems with map completion.

Instrumentation for Automated Colony Processing

Norman G. Anderson
Large Scale Biology Corporation, Rockville, MD 20855
(301) 424-5989

The objective of this project is (1) to process large numbers of clones through colony picking, (2) to purify colonies by recloning as necessary, (3) to reposition the colonies in large rectangular arrays, (4) to process the DNA from each colony to allow the transfer of recombinant DNAs to the nitrocellulose or other filter materials, (5) to probe the arrays *en masse* with DNA probes under different conditions of stringency, and (6) to identify and order overlapping clones. We propose to develop an automatic cloning, repositioning, and hybridization analysis system based on proprietary technology that can ultimately handle sets as large as 2 million clones. The system is designed to do large matrix hybridization analyses at different and closely controlled stringencies and to recover the filter materials used for multiple reprobing. Key elements in the design are positive clone identification, arrangements for the production of duplicate sets of large arrays, and long-term storage of these arrays. The entire system is designed to run automatically in a remote environment under sterile conditions.

Abstracts: SBIR–Phase I

Increased Speed in DNA Sequencing by Utilizing LARIS To Localize Multiple Stable Isotope Labeled Fragments

Heinrich F. Arlinghaus
Atom Sciences, Inc., Oak Ridge, TN 37830
(615) 483-1113

The need to map and sequence the DNA in the human genome is of crucial importance to future medical advances and economic competitiveness. Utilization of DNA probes labeled with dozens of stable isotopes, instead of a single radioisotope, could increase the rate of sequence determination 100-fold or more, compared with current methods. The new approach described here—laser atomization resonance ionization spectroscopy (LARIS)—will be used to localize and quantify these isotopes in DNA fragments that had been separated by polyacrylamide gel electrophoresis. The unique selectivity and ultimate sensitivity made possible by the LARIS technique allow detection of isotope-tagged DNA fragments in the 10^{-18}-mol, or lower, range without isobaric or other interferences. This technique can be applied not only to genome sequencing but also to hybridization methods used in genetic engineering. For the proposed experiments in Phase I, we will use the same experimental apparatus used by the ongoing sputter-initiated RIS (SIRIS) DNA sequencing measurements. Doing this allows comparison and evaluation of these two techniques. During Phase II, recommendations will be made for the most fruitful directions in DNA sequencing development, and an RIS system dedicated to DNA sequencing will be specified.

PC Program for Automated Analysis of Stained DNA Gels

Jeffrey M. Stiegman
BioPhotonics Corporation, Ann Arbor, MI 48106
(313) 426-8299

Electronic image processing of one-dimensional (1-D) or two-dimensional (2-D) electrophoretic gels requires specialized software for analysis. Although several commercially available packages provide generic image enhancement and spatial analysis, a high degree of user modification is necessary to produce useful data for 1-D or 2-D gels. Dedicated software for electrophoretic gels has focused primarily on 2-D protein separation. However, the application of electronic image analysis to 1-D separations of DNA fragments has received relatively little attention.

The goal of the BioPhotonics Corporation project is to develop computer software for capturing, enhancing, and documenting electronically acquired images of 1-D electrophoretic separations of fluorescently stained DNA fragments and for subsequently converting this data to molecular weight and concentration information to be used for statistical analysis and restriction fragment mapping. The emphasis of this project is to automate fully the procedures for locating band positions, determining relative DNA concentration, calculating molecular weights, and analyzing migration abnormalities. This approach will provide improvement in time and accuracy many times over that of manual methods.

Abstracts: SBIR–Phase I

Rapid Genome Analysis on a Workstation with an Associative Coprocessor

Charles D. Stormon, James Brule, and Hamid Bacha
Coherent Research, Inc., Syracuse, NY 13202
(315) 426-0929

The goal of this research project is to produce a high-performance, low-cost sequence analysis workstation that will allow genetic researchers to perform complex analyses on genome data. Phase I research involves feasibility studies of processing genome data at a workstation with an associative coprocessor. Many of the DNA sequence analyses that molecular biologists require involve finding similarities between different molecules or within one molecule. Several different approaches, represented by dozens of programs, have been taken to solve this problem. With increasing sequence length (or larger numbers of fragments) to be analyzed, these methods can quickly become computationally impractical. Recently, parallel processors have become available, and they provide an avenue of speedup for sequence analysis.

By implementing a simplified computational model in VLSI (very large scale integrated), our associative Coherent Processor™ (the AP-DS) provides the same degree of parallel processing as either the Connection Machine™ (Thinking Machines, Inc.) or the DAP™ (Active Memory Technology) at a much lower cost. The research in this project will show the feasibility of processing complex genome queries on a workstation-class machine with attached AP-DS coprocessor.

New Recording Media for an Automated Genome Mapping System

George M. Storti
Quantex Corporation, Rockville, MD 20850
(301) 258-2701

New recording media for an automated genome mapping system is the goal of this project. The media are different formulations of Quantex's electron trapping (ET™) materials that have the capability of recording beta particle events and events produced by visible light of relatively short wavelength. Consequently, the presence of radioactively tagged and dye-tagged DNA fragments can be recorded, and event information can be integrated. Readout will be performed by an X-Y scanning system that allows for large signal-to-noise ratios. Compatibility of a scanning system with large-area DNA gels will be evaluated.

Improvements in DNA Sequencing Methods by Incorporating Chemiluminescent Detection

John C. Voyta
Tropix, Inc., Bedford, MA 01730
(617) 271-0045

The research plan for Phase I of this project will involve feasibility studies of the use of chemiluminescence in DNA sequencing. With the advent of enzymatically induced chemiluminescence from stable precursors and the improved instrumentation for luminometry, ultrasensitive detection of a wide variety of agents under many different conditions is now possible. With the implementation of the chemiluminescent substrate technology, we believe that it will be possible to improve several aspects of DNA sequencing methodology such as stabilizing reagents, eliminating radioisotopic labels, shortening the incubation times, improving detection sensitivities, and greatly simplifying the detection instrumentation. To verify the viability of such a system, we plan to use an avidin/biotin detection system with alkaline phosphatase as the signal-generating label. In one study, biotinylated primer strands and other necessary reagents will be used in the dideoxy-sequencing reactions and separated via gel electrophoresis. Subsequently, the separated DNA will be visualized with avidin-labeled alkaline phosphatase and a chemiluminescent substrate. The use of chemiluminescent substrates in multiplex-sequencing protocols will also be investigated. In Phase I, the luminescent detection of nucleotides will be accomplished with X-ray and instant films. The anticipated improvements in chemiluminescence-based DNA sequencing will be fully explored in Phase II with the implementation of full-performance protocols coupled with the design and prototype generation of a simple and inexpensive luminescent gel/blot scanner capable of high throughput. Finally, we believe that these new and improved sequencers will offer speed, cost effectiveness, and ease of use in DNA sequence determinations.

Chemiluminescent Labels for Polynucleotides Electrophoretically Separated in Agarose Gels*

Edward M. Davis
SymBiotech, Inc., Cheshire, CT 06410
(203) 272-4190

The objective of this project is to continue the work, started during Phase I, which demonstrated the feasibility of employing chemiluminescence to visualize polynucleotides separated by electrophoresis in agarose gels. Chemiluminescence is a very powerful labeling method and, under optimal conditions, yields detection limits equal to or better than radioisotopic labels. Much Phase II work will be directed toward improving the detection limits for DNA in electrophoresis gels. This improvement will be accomplished by developing ways to decrease background luminescence and increase the efficiency of biotinylation of DNA, as well as to explore more powerful luminescent reagents. Ultimately, the aim is to develop highly sensitive and safe labels that can be detected with inexpensive photography or, alternatively, can be recorded, processed, and stored by an electronic digital imaging system. The potential of employing chemiluminescence for labeling DNA fragments in sequencing gels will also be explored. It is anticipated that several products will be developed from this work including a combination electrophoretic device for labeling and photographing luminous DNA. In addition, new reagents for improved luminescence and biotinylation of DNA will be developed for commercialization.

*1988 award for two years.

The Separation of Large DNA Fragments with Oscillating Electric and Magnetic Fields*

Gunter A. Hofmann
BTX/Biotechnologies and Experimental Research, Inc., San Diego, CA 92109
(619) 270-0861

The accurate and fast separation of large DNA fragments is a crucial technology needed for mapping the human genome. Existing methods such as pulsed-field gel electrophoresis appear to have severe limitations. This project uses a novel approach: Lorentz-force-mediated separation in the form of oscillating electric and magnetic fields. DNA exhibits a large induced dipole moment at low frequencies, with a relaxation time dependent on the length of the fragment. The planned separation method makes use of the polarizability of DNA fragments by subjecting them to an oscillating electric field with a superimposed perpendicular oscillating magnetic field in a liquid without matrix. The resulting unidirectional Lorentz force moves the DNA fragments perpendicular to both the electric and the magnetic field with a drift velocity dependent on the DNA polarizability, which depends in turn on the DNA size and configuration. Phase I studies demonstrated that DNA fragments can be polarized by an oscillating electric field and that a superimposed magnetic field results in a unidirectional Lorentz force and a unidirectional drift velocity. The drift velocity increases with the size of the DNA fragment, in contrast to conventional electrophoresis. For large DNA fragments, the drift velocity is two to three orders of magnitude higher than the drift velocity observed in experiments in pulsed-field gel electrophoresis. A theoretical model with a porous sphere simulation of the DNA molecule showed promise in yielding some qualitative agreements with the observed experimental dependencies of the drift velocity. Phase II studies will develop several prototype apparatuses of increasing complexity that will allow the rapid, accurate separation of large DNA fragments. A wider parameter range will be investigated, especially separation at higher frequencies, and optimum operating conditions will be determined. The limits of the crossed-field separation method will be investigated in experiments and theory.

•1989 award for two years.

An Image Acquisition and Processing System for the Analysis of Fluorescence from Stained DNA Gels*

Ronald A. McKean
KMS Fusion, Inc., Ann Arbor, MI 48106
(313) 769-8500

Current methods that use fluorescence techniques to analyze stained DNA gels are inadequate because the test results (1) cannot be easily accessed by a computer and (2) cannot be precisely reproduced. The lack of available instrumentation to convert a fluorescing image into a digitized record for computer entry and the lack of any technique for standardizing analyses performed under varying conditions severely limit the usefulness of this analysis. Overcoming these problems is essential as the need increases for an efficient means of performing both the analysis and the statistical review of the results. The goal of this project is to develop an instrument that will digitize stained DNA directly from agarose gels, standardize the results, and provide statistical analysis features. This instrument will quickly scan the gel with an optical system capable of high photometric resolution as well as very high spatial precision. The digitized data will then be electronically preprocessed (e.g., background subtraction, filtering, data compression) to ensure correct, noise-free data structured for storage with minimal memory requirements. The stored data, then available for use in imaging the gel, will produce statistical comparisons and generate graphical displays. The data will be archivable using media such as magnetic tapes or disks. Phase I efforts have been highly successful. The results demonstrated (1) acquisition of high-quality image data directly from agarose gels without elaborate or high-cost components, (2) feasibility of standardizing DNA results under varying electrophoretic conditions, and (3) preliminary optical and electronic designs for this instrument. Phase I results thus show the feasibility of the instrument that will solve many problems associated with agarose gel analysis of DNA.

*1988 award for two years.

A Novel, Low-Cost System for Automated DNA Sequence Reading*

John West
BioAutomation, Inc., Bridgeport, PA 19405
(215) 275-4540

Analytical instruments for the automated measurement of DNA are just emerging. Automated sequence readers commercialized to date are expensive (over $90,000 per instrument). As a result, in more than 10,000 sequencing laboratories worldwide, fewer than 3% have acquired such instrumentation. Because of the high cost, adoption of automated DNA-sequencing instrumentation has been slow even though interest in the technology is high. First-generation sequence readers also lack the throughput required for large-scale sequencing projects. Technological feasibility of an autoradiography-based approach incorporating the use of application-specific, highly integrated circuits and parallel processing was demonstrated in Phase I. During Phase II, full working prototypes will be developed using this approach. These prototypes will (1) provide a significant increase in accessibility of automated sequence-reading equipment to smaller laboratories, and (2) provide the quantum jump in throughput required by large-scale sequencing projects.

Being proprietary, the specific technical approach to be followed is omitted from this abstract.

*1989 award for two years.

Index to Project Investigators

Abstracts

Abstracts: Index to Project Investigators

Resource Development

DNA sequence data output. Genetically modified T7 DNA polymerase lacking 3′ to 5′ exonuclease activity was used in the presence of manganese ions (Mn^{2+}) to catalyze the four separate polymerization reactions prior to sequencing the vector, M13 mp18 DNA, on an Applied Biosystems Model 370A Automated Sequencing System. Each polymerization reaction was carried out with a primer labeled with a different fluorophore and a different dideoxynucleoside triphosphate chain terminator [Smith et al., *Nature* **321**, 674 (1986)]. After completing the polymerization reactions, the mixtures were combined and run in a single lane on a denaturing polyacrylamide gel. The fluorescently labeled bands separated during electrophoresis were analyzed using the Applied Biosystems 370A. The red, blue, green, and black peaks correspond to ddTMP-, ddCMP-, ddAMP-, and ddGMP-terminated fragments, respectively. Shown is a part of three consecutive panels of a 450-nucleotide region of M13 mp18 DNA; the sequence of the first 20 nucleotides is TCGTACTCTAGAGGATCCCC. The height of the signal decreases with the length of the fragment, because statistically there are fewer terminations at each position with increasing length of the DNA strand being sequenced.

Because native T7 DNA polymerase has an inherent 3′ to 5′ exonuclease activity that causes premature terminations of polymerizations at pause sites, T7 DNA polymerase has been modified genetically so that it no longer has 3′ to 5′ exonuclease activity. However, when the genetically modified T7 DNA polymerase discriminated against the dideoxynucleoside triphosphates, uneven band intensities resulted when the usual polymerization reaction component—magnesium (Mg^{2+})—was present. Substituting Mn^{2+} in the polymerization reaction alleviated the problem of nonuniform DNA band intensity [Tabor and Richardson, *Proc. Natl. Acad. Sci. USA* **86**, 4076 (1989)]. (Photograph provided by S. Tabor and C. C. Richardson, Harvard Medical School.)

Appendices

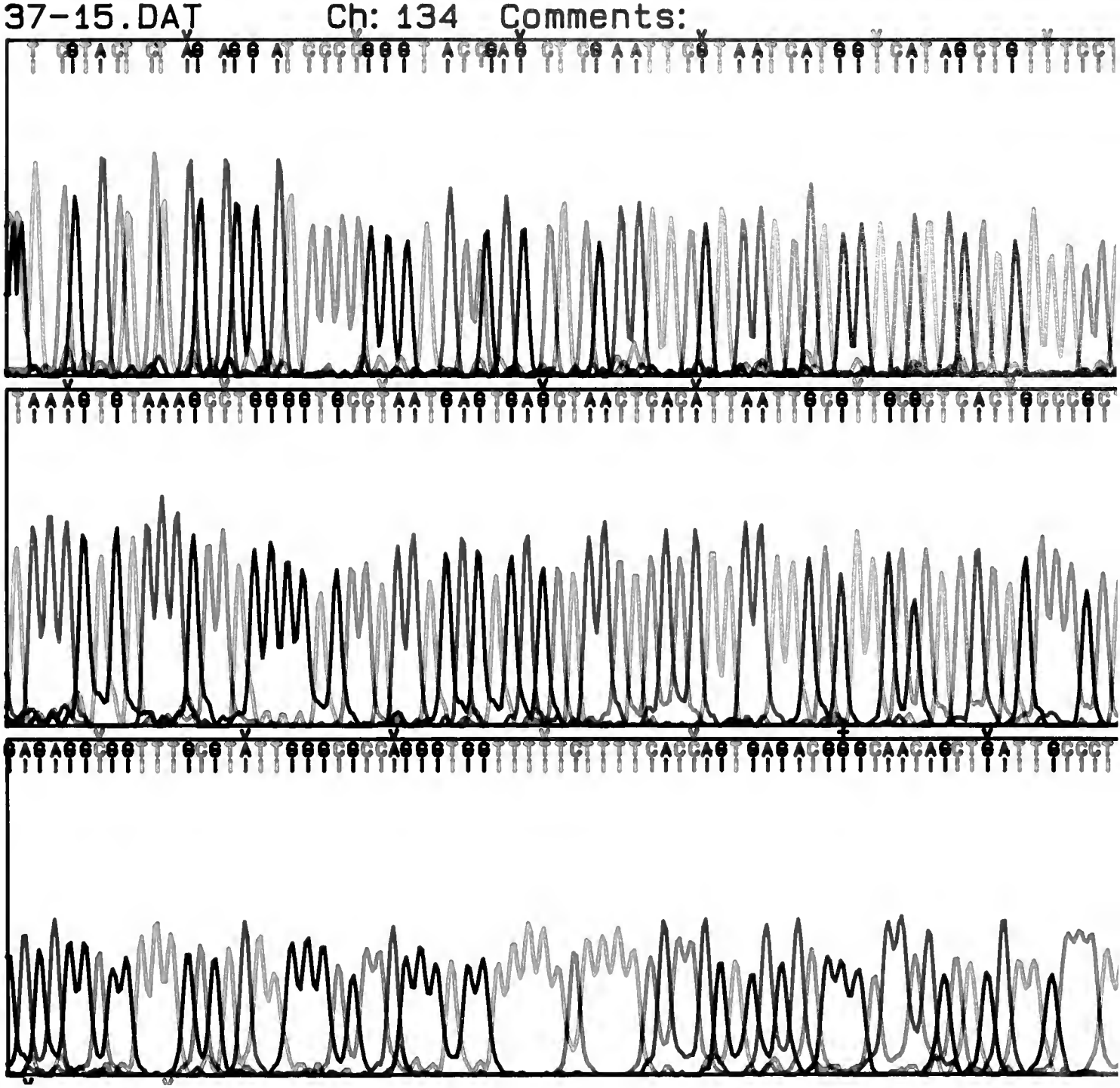
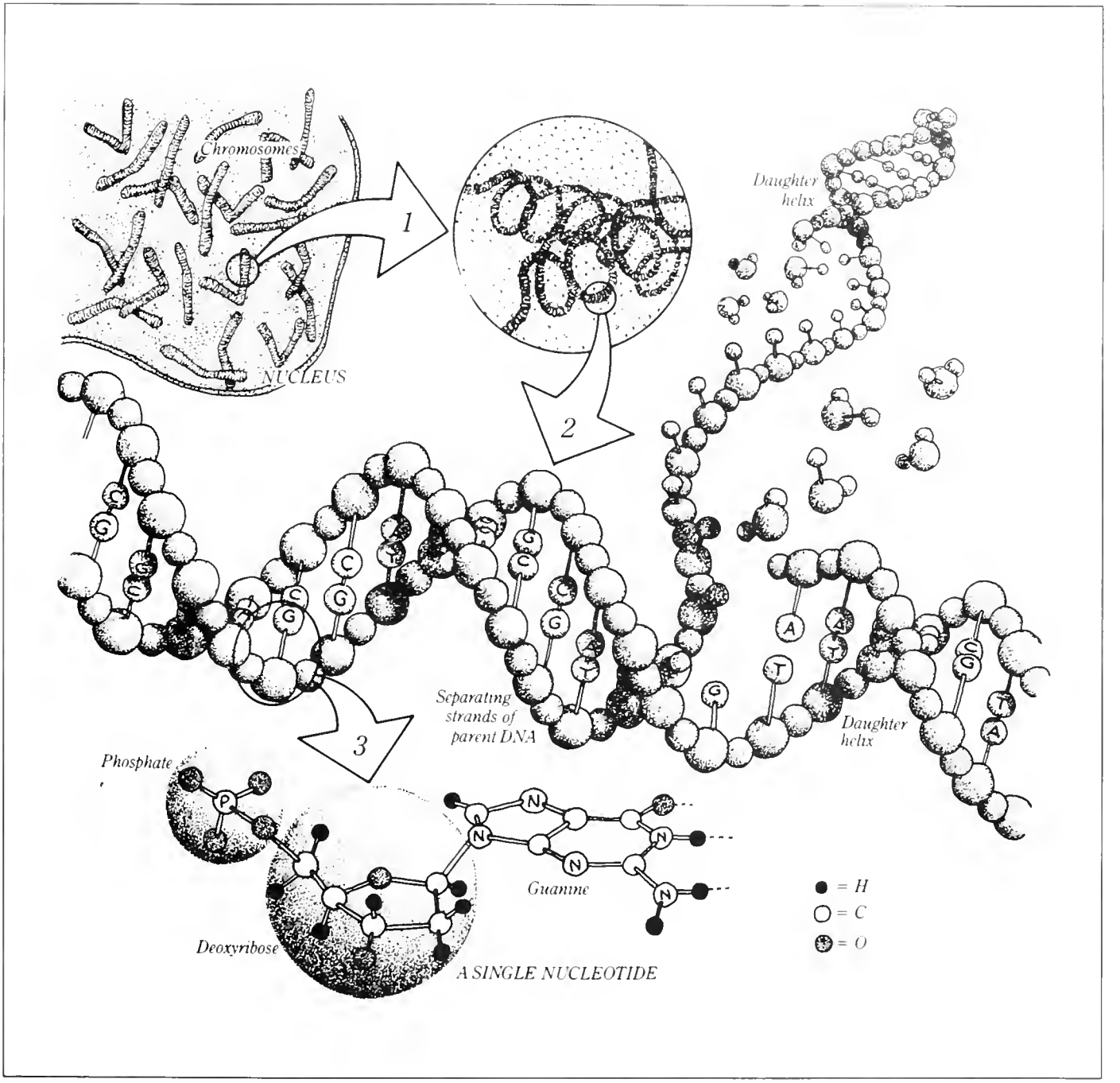


Fig. 1. The human genome at four levels of detail. Apart from reproductive cells (gametes) and mature red blood cells, each cell of the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA. Each strand of the DNA is a huge natural polymer consisting of repeating nucleotide units—each of which is comprised of a phosphate group, a sugar (deoxyribose), and a base (either guanine, cytosine, thymine, or adenine). In its "normal" state, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between guanine and cytosine and between thymine and adenine. Each such linkage is said to constitute a "base pair"; some 3 billion bp constitute the human genome. The specificity of these base-pair linkages underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand; the nucleotide sequence (i.e., linear order along the DNA strand) of each strand is strictly determined. Each daughter double helix is thus not only a twin, but also an exact replica of its sole parent. (Figure and caption text provided by the Human Genome Center, Lawrence Berkeley Laboratory.)

Primer on Molecular Genetics

Appendix A



Appendix A: Primer on Molecular Genetics

The text for this primer was adapted from material and figures that were supplied by Charles Cantor and Sylvia Spengler of the Human Genome Center, Lawrence Berkeley Laboratory.

Introduction to the Genome

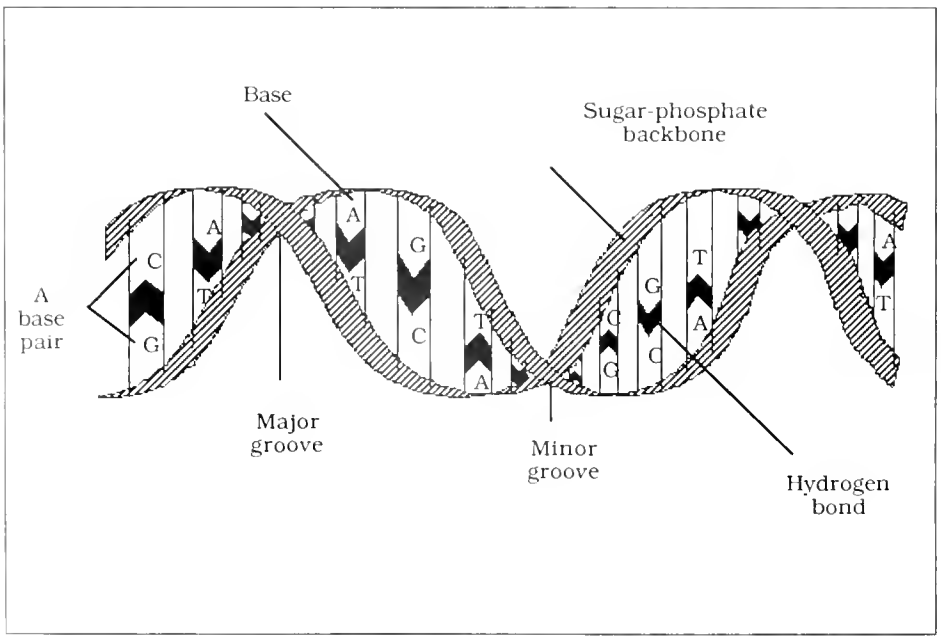
The genome is the total complement of genetic material present in a single cell. This genetic material is deoxyribonucleic acid—DNA—the blueprint for all cellular activities for the lifetime of the cell or organism. Each of a person's 10 trillion cells contains essentially the same complement of DNA (except for mature red blood cells, which contain no chromosomal DNA).

Chromosomes

The human genome, containing 3 billion base pairs of DNA, is divided into 23 pairs of physically separate units called chromosomes. Chromosomes, which are located in cellular nuclei (Fig. 1, p. 121), contain roughly equal parts of protein and DNA. Each chromosome has a single DNA molecule (packaged in a complex hierarchy and whose number of bases average 150 million) that would be 2 in. long if released from the cell and stretched out. DNA molecules are the largest molecules now known.

Chromosomes can be seen under the light microscope. Cytological stains reveal a pattern of light and dark bands that reflect local variations in the percentage of Adenine and Thymine versus that of Guanine and Cytosine in particular regions. Differences in size and banding pattern allow each of the 24 chromosomes to be distinguished from the other. With microscopic examination, one can detect occasional major chromosomal abnormalities that indicate differences in the genomes of some individuals. However, the majority of DNA differences are more subtle and can be detected only by molecular analysis. These abnormalities in DNA may be responsible for inherited diseases or cancer.

Fig. 2. A diagram of a very short section of the human genome, which contains about 200 million times the amount of DNA shown in this section. [U.S. Congress, Office of Technology Assessment, *Mapping Our Genes—The Genome Projects: How Big, How Fast?* OTA-BA-373 (Washington, D.C.: U.S. Government Printing Office, April 1988).]



DNA

The DNA structure consists of a long, ribbon-like chain made of two strands and divided into subunits. Each position along a DNA strand can be occupied by one of four nitrogenous bases—adenine (A), thymine (T), cytosine (C), or guanine (G)—strung together along a backbone of sugars and phosphates (Figs. 1 and 2). The sequence of bases along the sugar-phosphate backbone encodes the

genetic information. The two DNA strands are related by the following rules for matching—pairing—the four bases:

- A on one strand is always matched by T on the other to form an A-T base pair.
- G on one strand is always matched by C on the other to form a G-C base pair.

Adherence to these rules ensures that a duplicate information set is available to correct errors and thus minimize the mutation rate.

When a cell divides, its DNA divides too. Using the rules for forming base pairs, each strand directs the synthesis of a complementary new strand. Each daughter cell receives one old and one new DNA strand (Figs. 1 and 3).

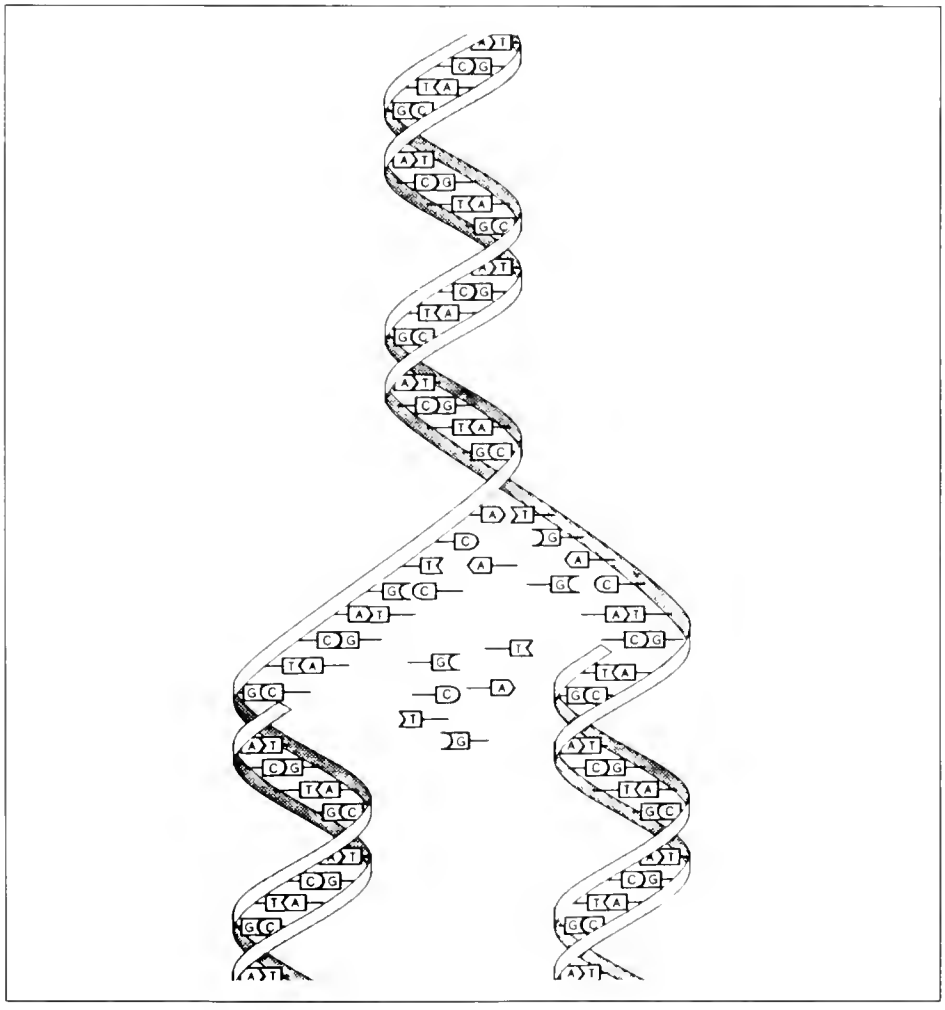


Fig. 3. Replication of DNA. Complementary DNA strands separate; each strand then becomes the template for synthesis of a new strand.

Appendix A: Primer on Molecular Genetics

Genes

Genes—the functional unit of genetic information—encode the DNA sequences required to make a protein. Genes also contain information that is used to regulate the kind and amount of protein made in a particular type of cell.

The human genome has 50,000 to 100,000 genes. A typical gene is 90% stuffer—introns—whose function is unknown. A human gene may contain up to 30,000 base pairs, but only about 10% of the base pairs are known, with certainty, to contain useful information. This information is read out in discontinuous blocks, called exons (Fig. 4). Since three bases code for one amino acid (the building blocks of proteins), the average size of the protein coded by a gene will be 1000 amino acids. In addition to introns and sequences that regulate the amount of protein made, special regions within the DNA help organize it into chromosomes and control its replication.

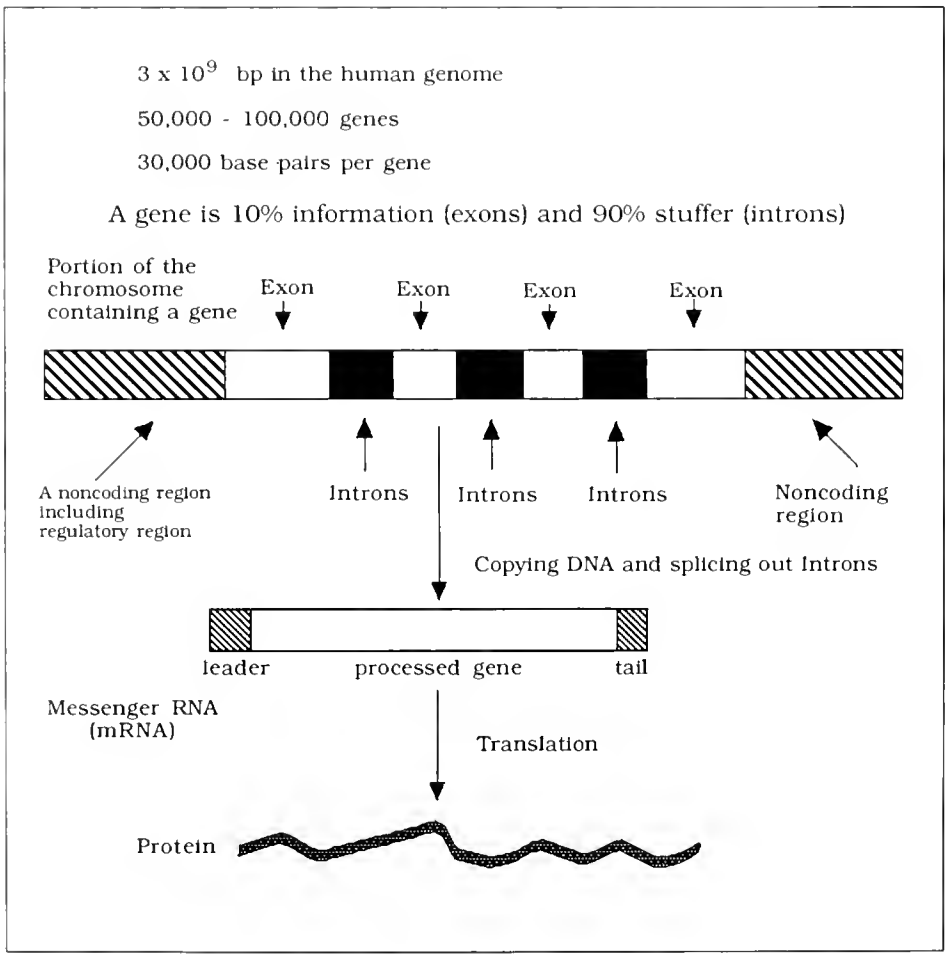


Fig. 4. Structure of human genes. When a gene is turned on by signals at the regulatory region, the entire gene—introns and exons—is copied. Then the introns are spliced out to make the shorter mRNA that is translated into protein by the cell.

Mapping and Sequencing the Genome

To determine the order of billions of pairs of raw DNA sequence, first the human genome must be broken down into genes or other fragments small enough to be propagated by cloning techniques and characterized for identification. Next, these chromosomal fragments will be ordered—mapped—to correspond to their respective locations on the chromosomes. Finally, automated techniques will be utilized to determine the base sequence of the ordered fragments—the ultimate goal of the Human Genome Project.

Molecular Techniques Used in the Human Genome Project

Improvements in mapping and sequencing techniques are a major focus of the genome project. The efforts will include developing automated probe and mapping methods, as well as optimizing techniques to extract the maximum useful information from maps and sequences. Some of the techniques currently being employed are described below.

Using Restriction Enzymes

Isolated from various bacteria, restriction enzymes serve as microscopic scalpels that recognize short sequences of DNA and cut the DNA molecules at those specific recognition sites (Fig. 5).

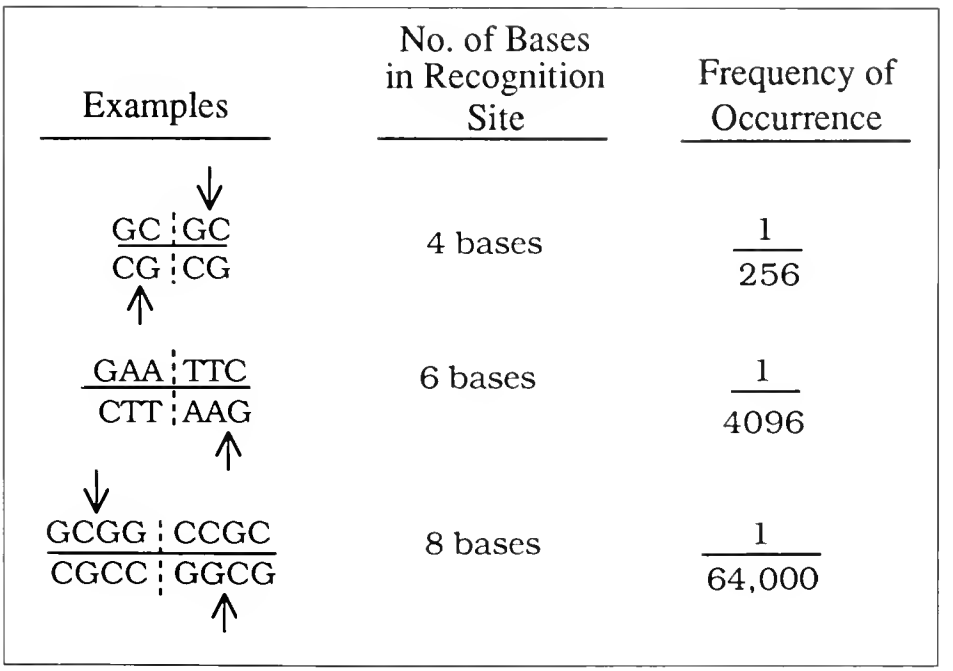


Fig. 5. Typical restriction enzyme cutting sites. The recognition site has a twofold symmetry around a point (the dashed line). The cutting sites (arrows) are on both strands of the double helix (above and below the solid line) and, when cut, generate smaller, linear fragments of DNA.

Appendix A: Primer on Molecular Genetics

On the average:

- 4-base recognition sites will yield pieces 256 bases long,
- 6-base recognition sites will yield pieces 4,000 bases long, and
- 8-base recognition sites will yield pieces 64,000 bases long.

Since hundreds of different restriction enzymes have been characterized, DNA can be cut into many different types of smaller fragments.

Electrophoretic techniques (Fig. 6) can be used to separate DNA fragments according to length with a 1-base-pair resolution up to 1,000 base pairs, with 1% resolution up to 1 million base pairs, and with 5% resolution up to 10 million base pairs. This great separation power forms the basis of most methods for handling and analyzing DNA molecules.

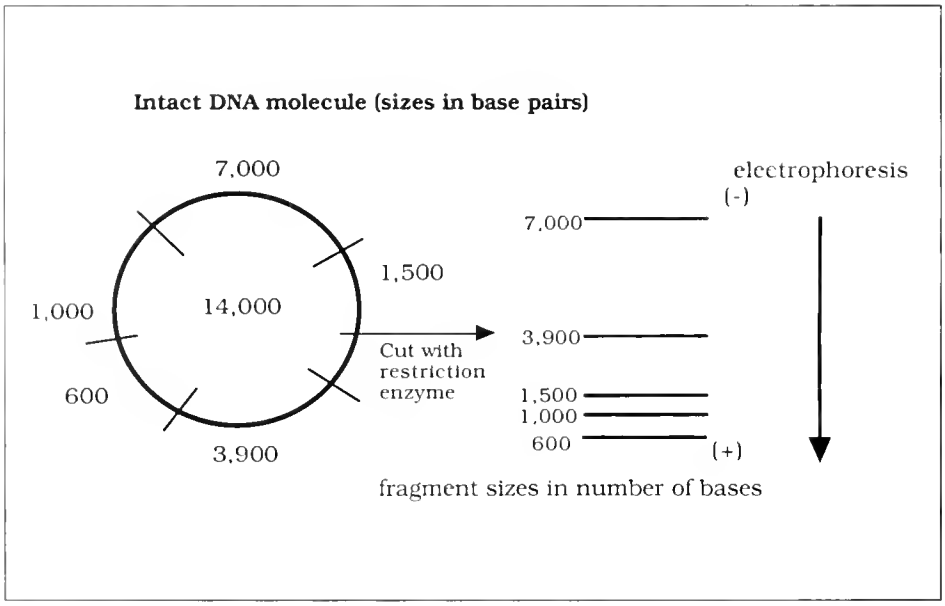


Fig. 6. Schematic results of electrophoresis of DNA. The fragments resulting from restriction enzyme cutting can be separated by gel electrophoresis. The smallest fragments move most rapidly toward the positive electrode. DNA fragments can be detected by uv absorbance, fluorescence, radioactivity, or other labels. The electrophoresis pattern itself can be of interest, since variations in the pattern from a given chromosomal region can sometimes be associated with variations in genetic traits, including susceptibility to certain diseases. This kind of "genetic linkage" has been used to establish the locations of important genes.

Cloning DNA Fragments

In vivo cloning. By propagating the fragments inside living cells, DNA pieces can be immortalized, and the amount of specific DNA can be amplified (Fig. 7). Bacteria are most often the hosts, but yeast and mammalian cells are also used. These cloning procedures provide unlimited amounts of material for experimental study.

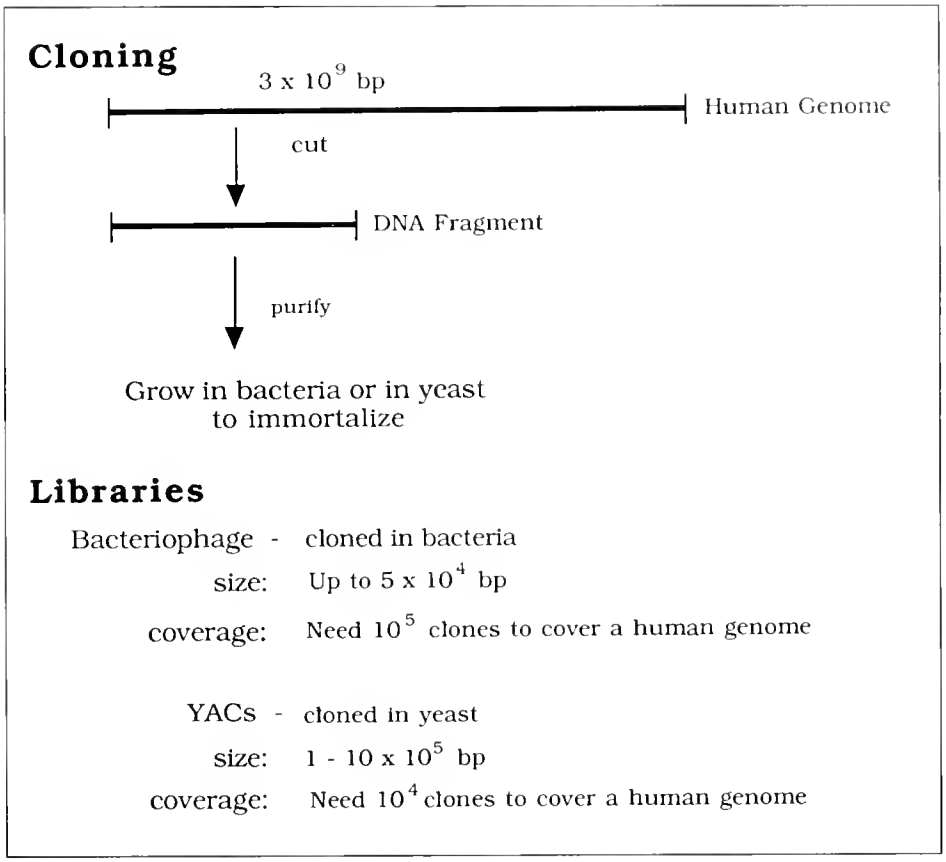


Fig. 7. Cloning and libraries. A DNA fragment can be cloned in bacteriophages or yeast artificial chromosomes (YACs). The minimum number of separate fragment clones needed in a library to cover the human genome of 3×10^9 bp is shown.

Appendix A: Primer on Molecular Genetics

In vitro cloning. A new approach to cloning, the polymerase chain reaction (PCR) amplifies DNA molecules in the test tube without using a host cell (Fig. 8). PCR is especially valuable because the reaction is easily automated and can handle samples too small or too toxic to be managed inside cells.

If a set of DNA fragments contains a complete sample of an entire genome or chromosome, it is called a library (Figs. 7 and 9).

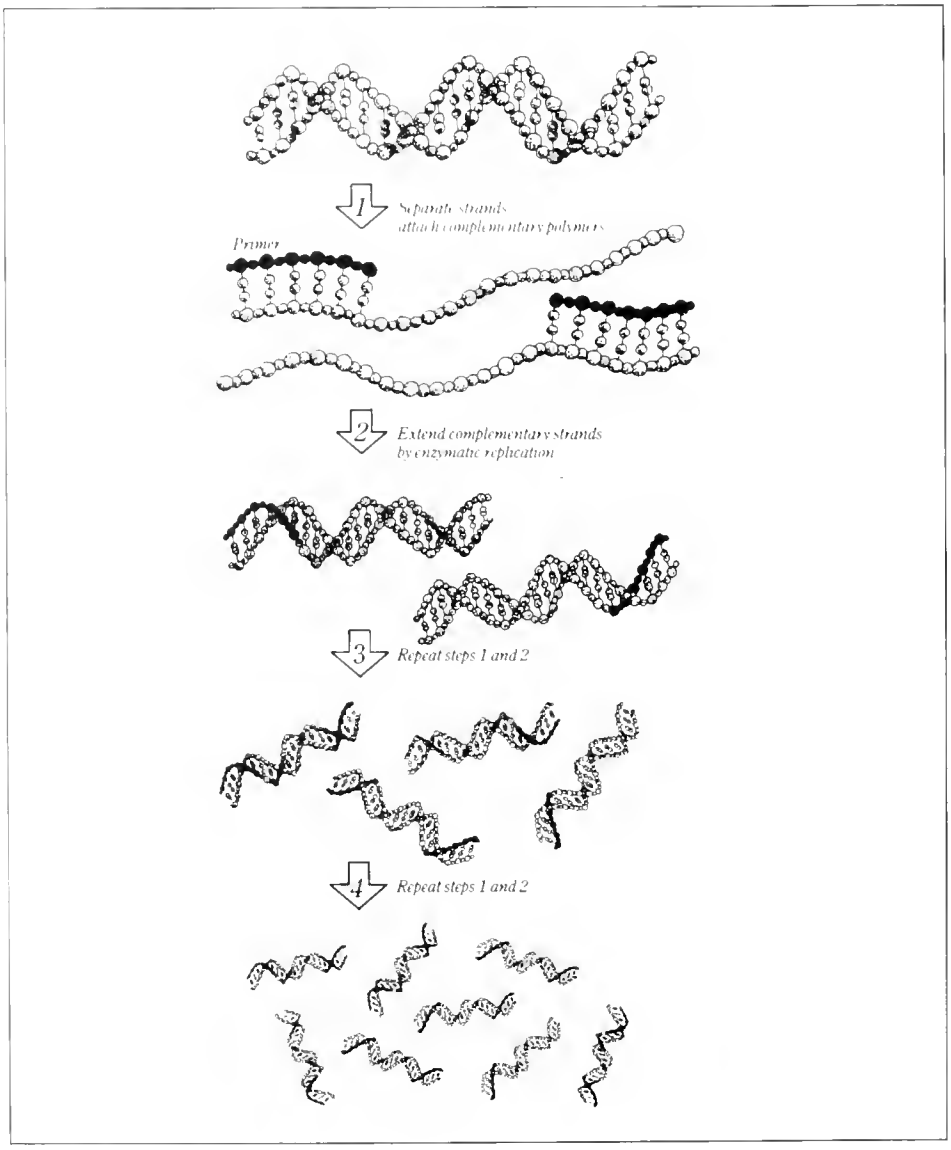


Fig. 8. DNA amplification by the polymerase chain reaction (PCR). PCR is a recently developed method for cloning human DNA—a necessary prelude to any intensive mapping or sequencing effort. PCR begins with the annealing of primers to the single-stranded fragments that are to be replicated. With the primers attached, a DNA replication enzyme can complete the complementary strands thus started. The succeeding PCR cycle then operates on the newly produced duplicates, as well as the original fragments. Each cycle thus doubles the amount of the selected DNA fragments in the reaction mixture. Twenty-five cycles, which can be completed in less than three hours, can theoretically produce amplification by 30 million-fold; in practice, amplification by a factor of over one hundred thousand can be readily achieved.

Mapping the Genome

The human genome may be "understood" at several levels (Fig. 10). Geneticists have already charted the approximate positions of over 1000 genes, and a patchy start has been made at establishing high-resolution maps of the genome. The coarsest maps are chromosome and genetic maps, which define the chromosomal locations of genes. The physical map is an ordered set of DNA fragments made from restriction enzyme fragments. The ultimate map is the base pair sequence for the human genome.

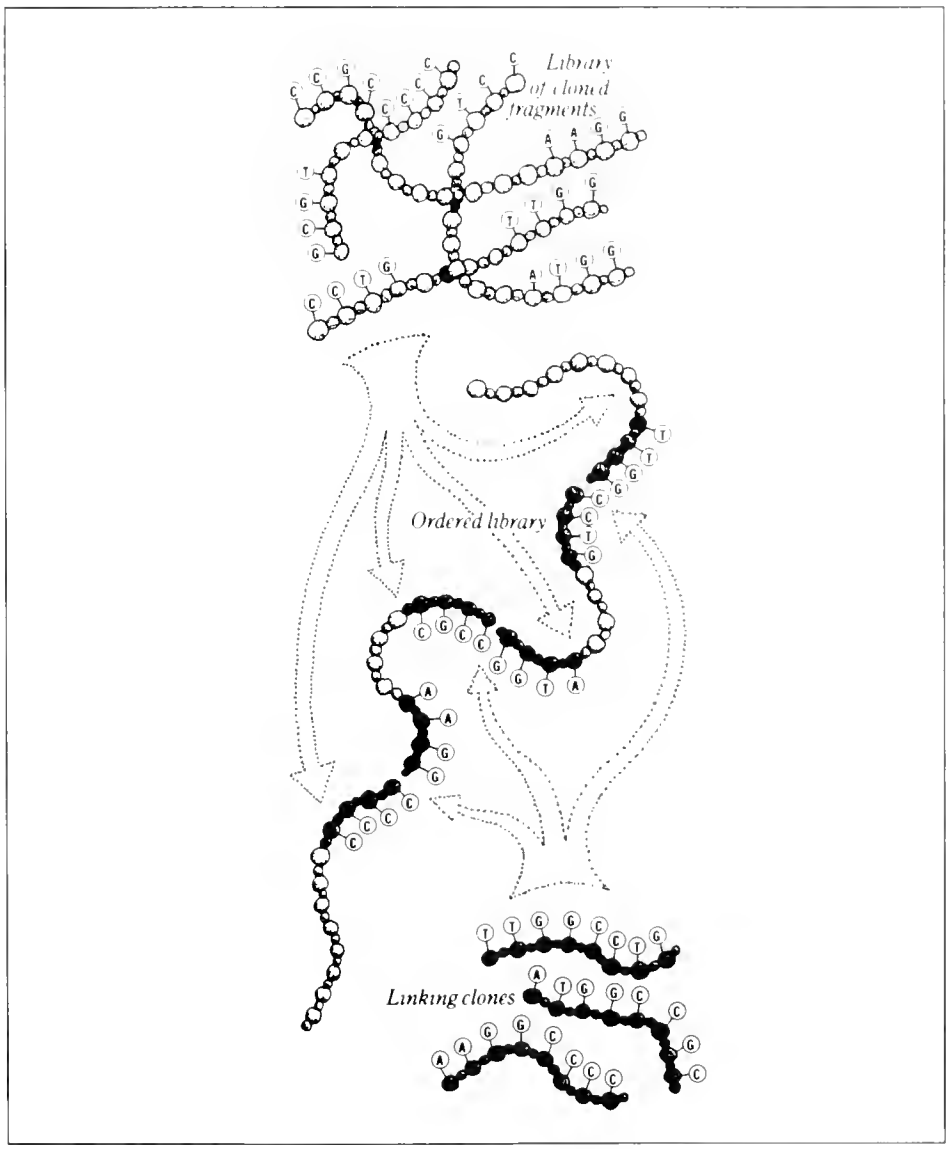
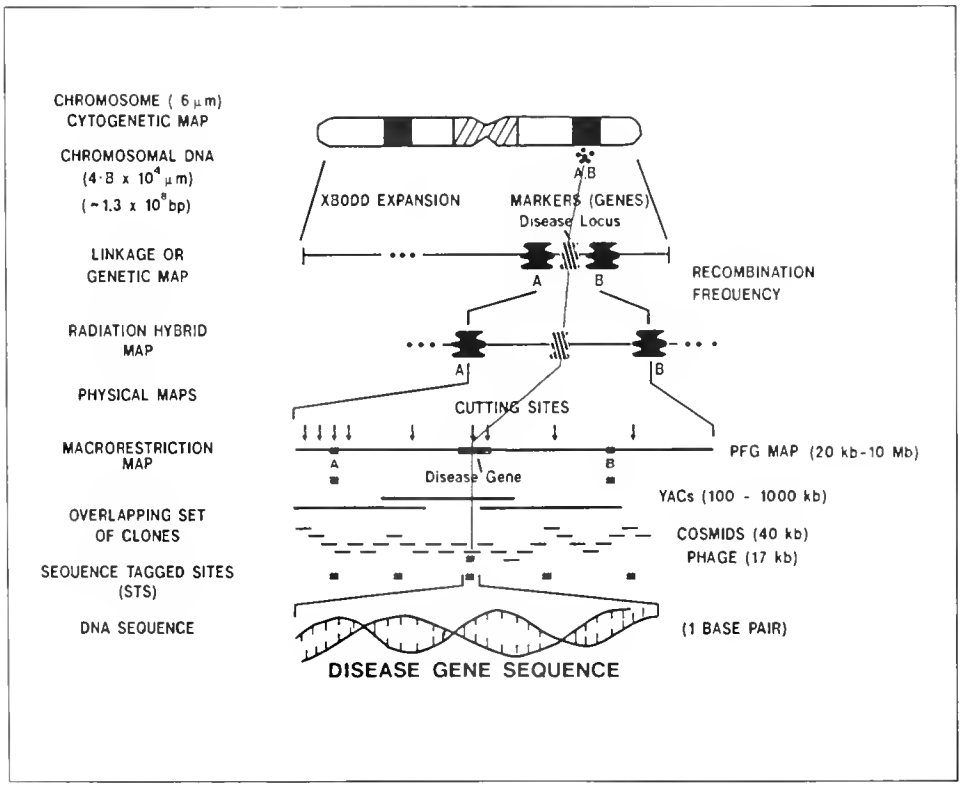


Fig. 9. Linking a library of cloned DNA fragments. Several strategies are available for ordering an unordered library of cloned restriction fragments. The one schematically illustrated here involves sequencing the ends of the fragments, together with a corresponding family of much shorter linking clones. The linking clones, themselves generated by restriction enzymes, are fragments that contain the cutting site (in this simplified example, GGCC) for the enzyme used to produce the clone library. In practice, linking clones are typically 50 to 100 base pairs long, and the fragments being linked are likely to be many thousands of base pairs long. If the library contains contiguous fragments and if the family of linkers is complete, assigning an order to the fragments is a simple matter of sequence matching—a puzzle-solving task best done by computer.

Appendix A: Primer on Molecular Genetics

Fig. 10. Multiple levels of mapping of human chromosomes. At the lowest level of resolution, the human genome is characterized by 22 pairs of autosomal chromosomes and two sex chromosomes—X and Y. Individual chromosomes can be distinguished by size, position of centromeric constriction relative to telomeres (centromeric index), and patterns of banding induced by staining with DNA-specific dyes after treatments that partially remove chromosomal proteins (G- and Q-banding). Cloned DNA probes can be used to determine by in situ hybridization the locations of specific genes or markers (A and B in this diagram) along a chromosome. To establish the order of—and relative distances between—genes or markers along specific human chromosomes, polymorphic DNA markers or fragments of gene regions are used to follow the inheritance of specific alleles in multigeneration families with large numbers of siblings. Distances on genetic linkage maps are given as percent recombination between markers: a distance of 1% recombination [1 centiMorgan (cM)] is generally estimated to represent approximately 1 Mbp of DNA. Radiation hybrid techniques can be used to obtain a higher-resolution (hundreds to thousands of kilobase pairs) physical map [S. Goss and H. Harris, *Nature* **255**, 1445–1458 (1975) and D. Cox et al., *Genomics* **4**, 397–407 (1988)]. The next level of resolution can be represented by the analytical physical—or macrorestriction—map that is developed using rare-cutting restriction enzymes and pulsed-field gel electrophoresis to separate megabase-sized DNA fragments. Markers are located on macrorestriction maps relative to restriction sites, and distances can be measured in kilobase pairs.



As an example of the value of long-range restriction maps, distances can be determined between markers flanking a disease locus. Overlapping sets of cloned DNA from phage or cosmid vectors and from yeast chromosomes—all engineered to accept foreign DNA—can be used for constructing these macrorestriction maps. A clone subset provides the substrate for obtaining the DNA sequence—the ultimate physical map. Maps from the cytological level to the DNA sequence are retained in various databases. The proposal regarding sequence-tagged sites (STSs) has offered that evenly spaced STSs would be used as identifiers to interrelate different levels of maps and to diminish requirements for storage of large numbers of biological reference materials [M. Olson et al., *Science* **245**, 1434–1435 (1989)]. (Figure provided by C. E. Hildebrand, Los Alamos National Laboratory.)

Chromosome maps

In a chromosome map, genes or DNA fragments are assigned to their respective chromosomes (Fig. 10). With in situ hybridization, the DNA (tagged with either a fluorescent or radioactive label) is used to seek out and bind to its complementary strand in an intact metaphase chromosome.

Until recently, the best chromosome maps could be used to locate a DNA fragment only to within a region of about 10 million base pairs—the size of a typical chromosome band. However, new methods promise to improve the resolution of the maps and allow narrowing of the range to around 1 million base pairs.

Genetic maps

Any molecular or physical characteristic that differs between individuals and is inherited is a potential genetic marker. The human genetic map is constructed by observing the pattern of inheritance of pairs of genetic markers. DNA sequence differences are especially useful markers because they are plentiful and easy to characterize precisely. Two markers located near each other on the same chromosome will tend to be passed together from parent to child (Fig. 11).

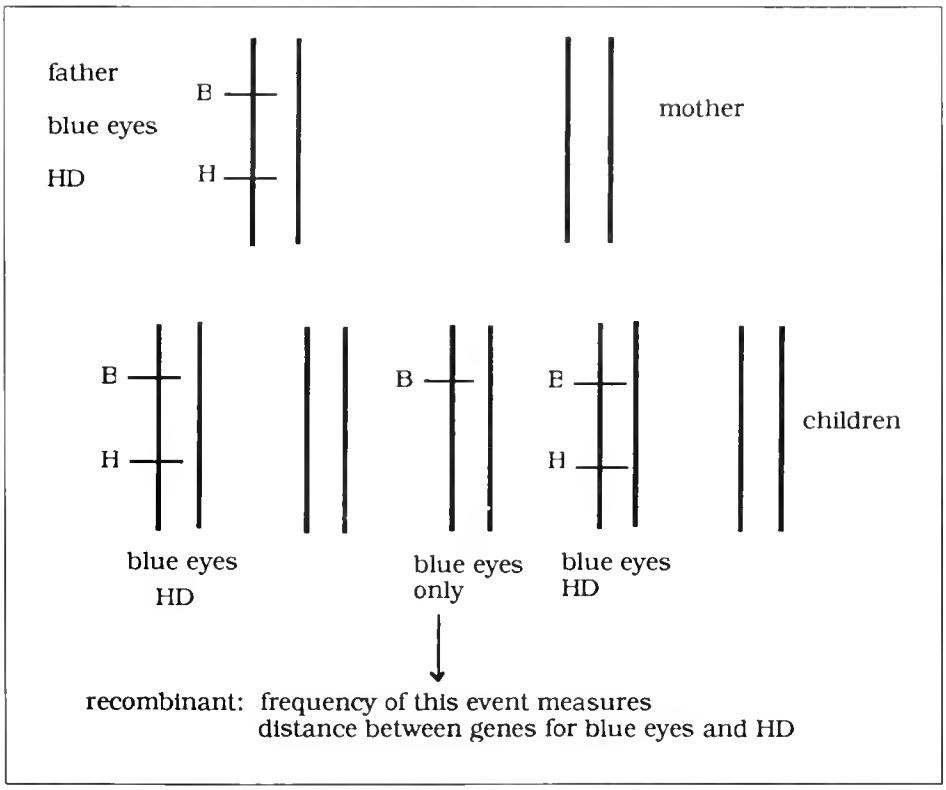


Fig. 11. How genetic maps are made. Each parent contributes one of each type of chromosome through the haploid egg and sperm so that the offspring will have a diploid chromosome set. In this illustration, the vertical lines show just one pair of chromosomes for each individual in a family. The father has, in this example, two traits that will be visible in any child that inherits them: blue eyes (B) and Huntington's disease (HD). The fact that one child received only a single trait, B, indicates that one of the father's chromosomes rearranged in the process of producing the child. (Note: In this example the mother's eyes were also blue.)

Appendix A: Primer on Molecular Genetics

Meiotic recombination, a relatively rare event, produces DNA strand breakage and rejoining; the result is separation of two markers originally on the same chromosome. The frequency of occurrence of these rare events provides an estimate of the distance between two markers. The closer they are, the less likely that a recombination event will fall between them and separate them.

The power of the genetic map is that an inherited disease can be located on the map, even though the molecular basis of the disease is not yet understood and the gene for the disease is not yet identified. Thus, genetic maps can be used to locate the neighborhood of important disease genes. The current resolution of the human genetic map is 10 million base pairs, about the same as the chromosome map.

Physical maps

Physical maps are ordered sets of DNA fragments made from restriction enzyme fragments. This mapping is done by either the top-down or the bottom-up approach (Fig. 12). Both of these approaches require purifying and analyzing large numbers of DNA fragments.

Top-down mapping. In top-down mapping, a single chromosome is cut into large pieces, which are then ordered, subdivided, and mapped further. DNA pieces can be located in regions about 100,000 to 1 million base pairs in size.

Bottom-up mapping. In bottom-up mapping, the chromosome is cut into small pieces. These pieces—the resulting clones—are then placed in order according to structural

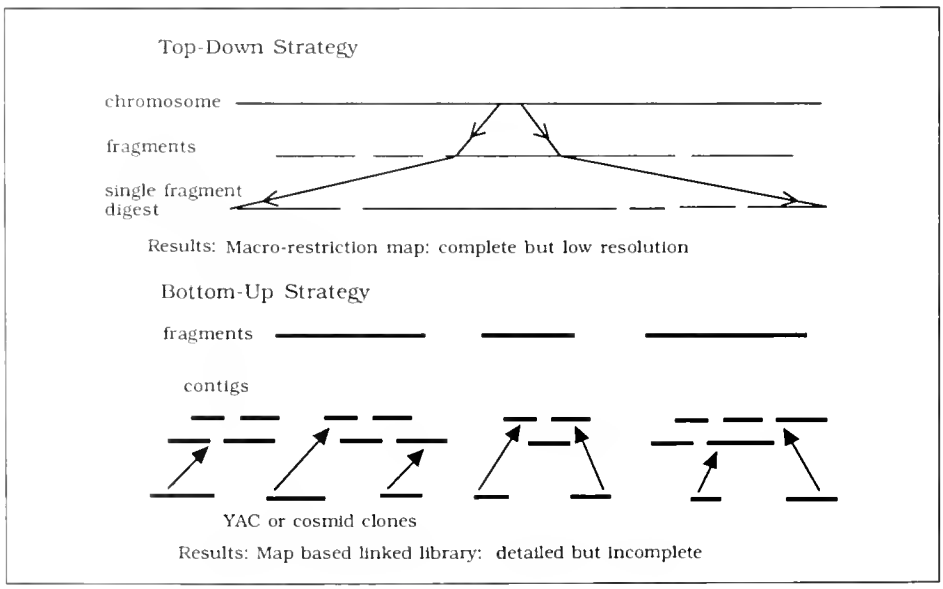


Fig. 12. Two physical mapping strategies. The top-down strategy gives low resolution but high connectivity. The bottom-up strategy has low connectivity because not enough clones can be mapped to give neighboring contigs.

features, such as restriction map data, that have two, and then multiple, clones in common. They then form contiguous blocks of DNA (contigs). The resulting ordered library of DNA pieces is 10,000 to 100,000 base pairs in size.

By using chromosomes purified by flow-sorters (a technique pioneered by the national laboratories) or in hybrid cell lines, one can concentrate on mapping a single chromosome at a time.

Sequencing

The DNA sequence is the ultimate physical map. Sequencing is also done by two basic approaches. Both of these methods work because very high resolution separations of DNA molecules are achievable with gel electrophoresis. In Maxam-Gilbert sequencing (Fig. 13A), DNA is cleaved at individual specific bases, and the lengths of the resulting fragments are determined. In Sanger sequencing (Fig. 13B), DNA replication is stopped at one of the four types of bases, and the lengths of the resulting DNA fragments are determined. Virtually all the steps in these sequencing methods are now automated.

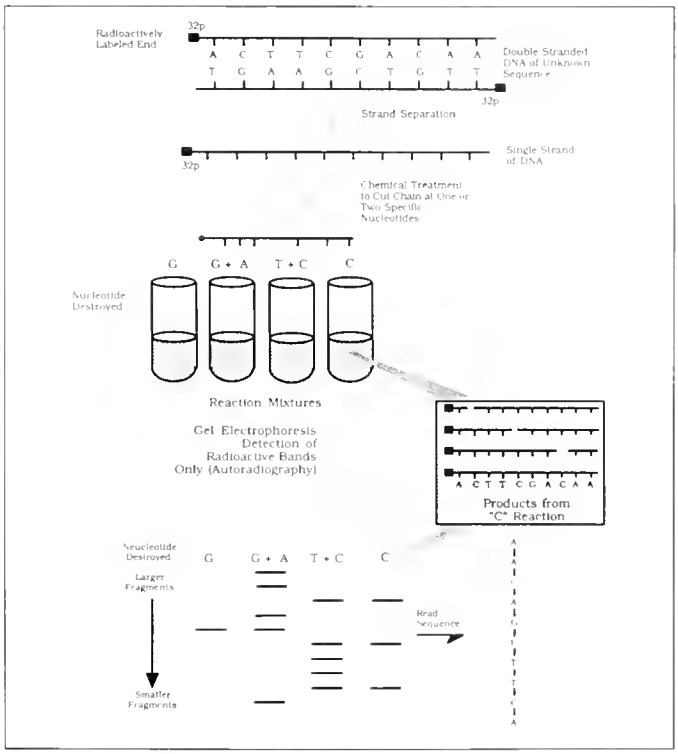


Fig. 13A. DNA sequencing by the Maxam–Gilbert method (Office of Technology Assessment, 1988).

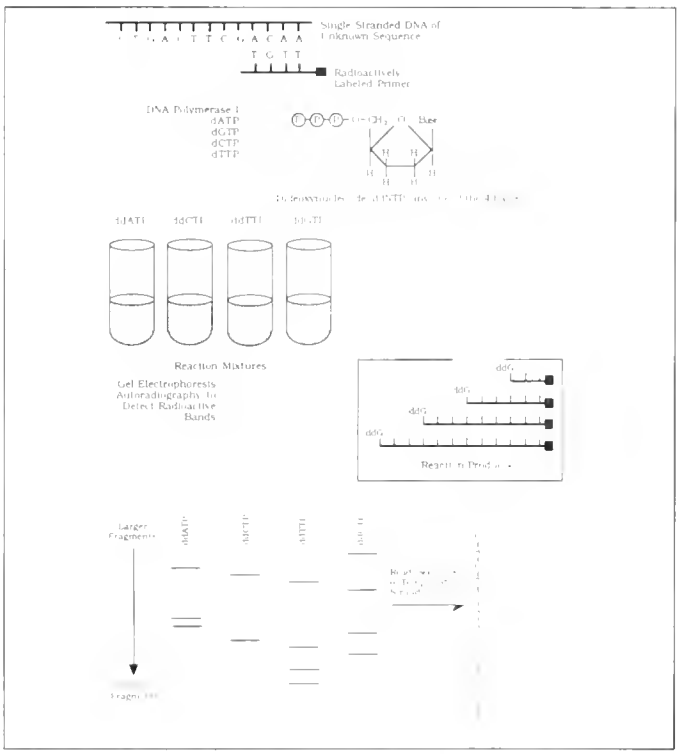


Fig. 13B. DNA sequencing by the Sanger method. (Office of Technology Assessment, 1988).

Appendix A: Primer on Molecular Genetics

End Games: Completing the Maps and Determining the Sequences

Starting maps and sequences is relatively easy; finishing them is very difficult. New methods are needed to expedite this end game.

Chromosome Walking with Primers

An approach being used to finish sequences is chromosome walking with primers (Fig. 14). In this technique, the walk begins with a primer at a known site and proceeds in a linear fashion—one step at a time—while adjacent regions that were previously unknown are being identified and sequenced. The larger region then serves as the new primer. All primers are synthesized chemically. The difficulty with this technique is the number of syntheses required for a large project.

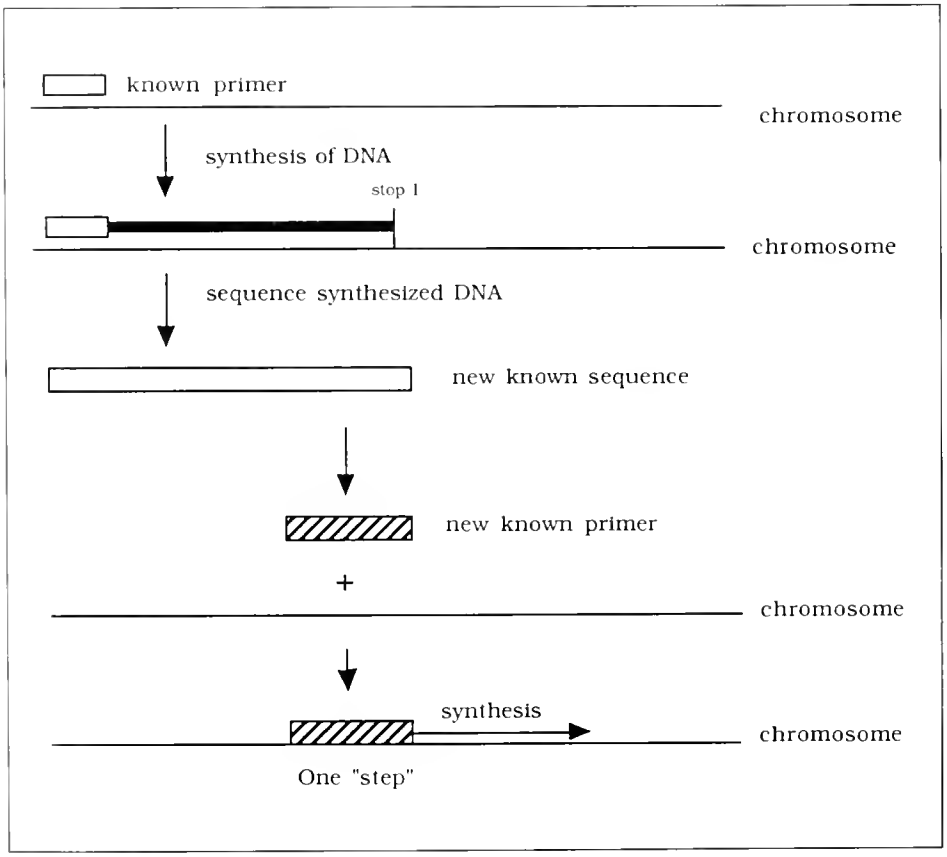


Fig. 14. Chromosome walking.

Single Chromosome Dissection

Another approach to finishing maps is to utilize single-chromosome dissection by physical methods (Fig. 15). This method isolates DNA pieces from the regions that are not yet mapped or sequenced and then applies the previously described mapping methods to those pieces.

Locating Specific Genes

The current human genetic map has about 1000 markers, or 1 marker spaced every 3 million base pairs. About 100 genes will lie between each pair of markers. In some regions of particular interest, genetic maps have been made that are five to ten times more detailed. By combining genetic and physical map information for a region, the stage is set for locating new genes (Fig. 16). The genetic map of these markers basically

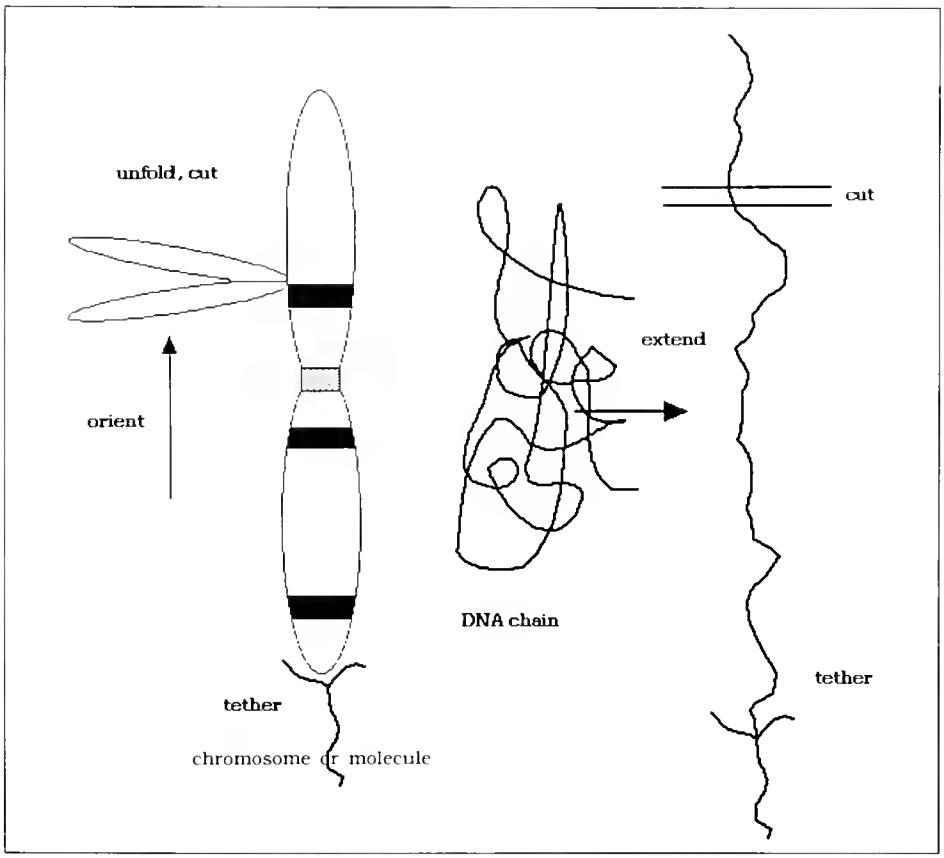


Fig. 15. Chromosome or single-molecule strategies.

gives gene order. Rough information about gene position is sometimes available also, but these data have to be used with caution, because recombination is not equally likely at all places on the chromosome; thus, the genetic map, compared to the physical map, stretches in some places and compresses in others. It is as though the genetic map were drawn on a rubber band.

How difficult it is to find an actual disease gene of interest depends largely on what is already known about that gene and, especially, on what sort of alterations in the gene have resulted in disease (Fig. 17). If disease results from a single altered DNA base, spotting the disease gene is very difficult; sickle cell anemia is an example of such a case, as are probably the majority of major human inherited diseases. When disease

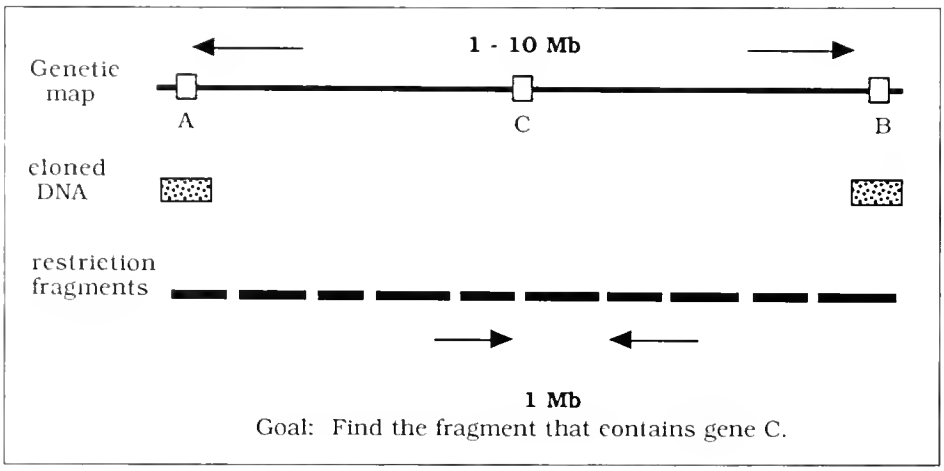


Fig. 16. Finding genes. A genetic map will reveal that a given gene of interest (C) lies in a region, perhaps encompassing 10 million base pairs (10 Mbp). The physical map allows one to dissect that region and to locate likely pieces on which the particular gene may reside.

results from a large DNA rearrangement, this anomaly can usually be detected by alterations in the physical map of the region or even by examination of the chromosome. The location of these alterations pinpoints the site of the gene.

To identify—without a map—the gene responsible for a specific disease, is analogous to finding a needle in a haystack. Finding the gene is even more difficult, because no matter how close one gets, the gene still looks like just another piece of hay. However, maps make finding genes much easier. They tell where to look in the haystack. The finer the map, the fewer the pieces of hay one has to test to see which is the gene of interest.

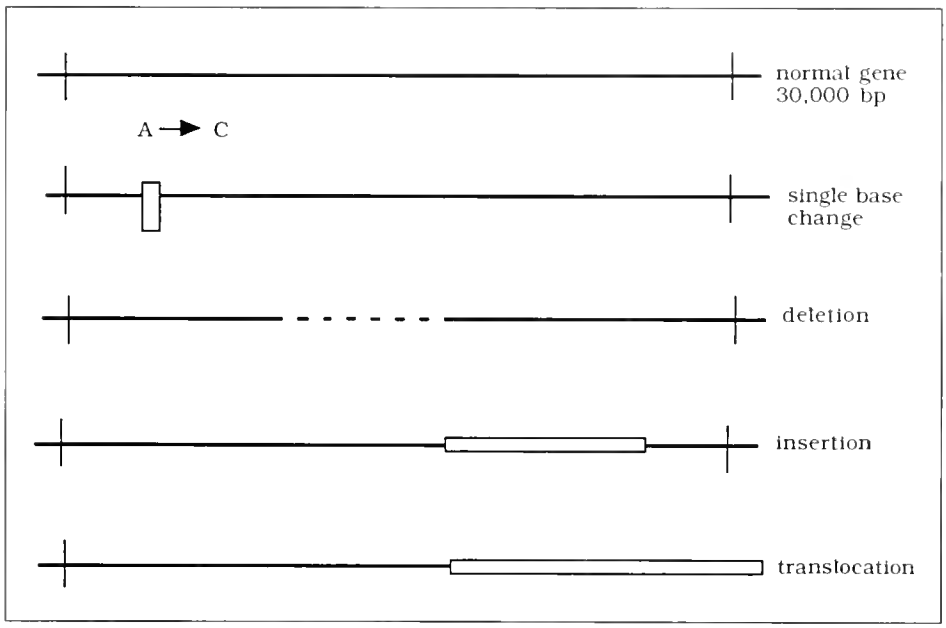


Fig. 17. Possible DNA abnormalities that can produce an inherited defect.

"The Alta Summit, December 1984"

Appendix B

Appendix C: "The Alta Summit, December 1984"*

Robert Mullan
Cook-Deegan

Alta is a ski area nestled among the Saguache Mountains in Utah, a winding 40-minute drive southeast from Salt Lake City. From December 9 to 13, 1984, visitors were isolated by repeated blizzards. The slopes were covered most mornings with Utah's renowned fine light powder, which beckoned skiers to cut its virgin surface.

For those 5 days, Alta was also a capital of human genetics. Many historical threads in the fabric that later became the Human Genome Project wind through that meeting, although it was not a meeting on mapping or sequencing the human genome. Through happenstance and historical accident, Alta links human genome projects to research on the effects of the atomic bombs dropped on Hiroshima and Nagasaki 40 years earlier. If genome projects prove important to biology, then historians will note the Alta meeting.

The Alta meeting was sponsored by the Department of Energy (DOE) and the International Commission for Protection Against Environmental Mutagens and Carcinogens. It was initiated by David Smith of DOE and Mortimer Mendelsohn of the Lawrence Livermore National Laboratory, who turned over final organization to Raymond White of the Howard Hughes Medical Institute at the University of Utah.

The purpose was to ask those working on the front lines of DNA analytical methods to address a specific technical question: could new methods permit direct detection of mutations, and more specifically could any increase in the mutation rate among survivors of the Hiroshima and Nagasaki bombings be detected (in them or in their children)? The idea behind the Alta meeting came from another meeting on March 4

TABLE 1

Participants at the Alta Meeting, December 1984

David Botstein	Mortimer Mendelsohn
Elbert Branscomb	John Mulvihill
Charles R. Cantor	Richard Myers
C. Thomas Caskey	James V. Neel
George Church	Maynard Olson
John D. Delahanty	David A. Smith
Charles Edington	Edwin Southern
Raymond Gesteland	Sherman Weissman
Michael Gough	Raymond L. White
Leonard Lerman	

*Reprinted with permission from Academic Press, Inc., *Genomics* **5**, 661–663 (October 1989).

and 5, 1984, in Hiroshima, at which new DNA analytical tools were deemed second highest priority for human mutations research, just behind establishing cell lines from atomic bomb survivors, their progeny, and controls. Those attending the Alta meeting in December (see Table 1) were drawn from a variety of backgrounds, and many had never met each other. Most said in interviews later that they came to the meeting quite skeptical, but left thinking it had been one of the best scientific meetings they ever attended (Interviews, 1987, 1988).

The principal conclusion of the meeting was, ironically, that methods were incapable of measuring mutations with sufficient sensitivity, unless an enormously large, complex, and expensive program were undertaken. Technical obstacles thus thwarted attainment of the main goal of the meeting, yet the meeting left a profusion of new ideas in its wake, some of which later washed ashore to be incorporated into various genome projects. Five years later, there is still no sensitive assay for human heritable mutations, but there *are* genome programs at NIH, at DOE, and in several foreign nations.

Excitement about the new methods blossomed at Alta despite, or perhaps because of, the wintry isolation. As Mortimer Mendelsohn noted in his internal report to DOE:

> It was clear from the outset that the ingredients for a successful meeting [were present]. . . and the result far exceeded expectation. Once the point of the exercise was clear to everyone, a remarkable atmosphere of cooperation and mutual creativity pervaded the meeting. Excitement was infectious and ideas flowed rapidly from every direction, with many ideas surviving to the end. (Mendelsohn, 1985).

John Mulvihill began the meeting by reviewing epidemiological studies of human mutations. Studies that could theoretically have detected a threefold increase in mutations had not found any. James Neel spoke about measurement of mutations among Hiroshima–Nagasaki survivors, estimating that the likely mutation rate was 10^{-8} per base pair per generation (or roughly 30 new mutations per genome per generation), indistinguishable from that of Japanese controls and in the same general range as that estimated by epidemiological methods and detection of protein variants among other "normal" populations. Several of the technical consultants commented on the passionate devotion Neel brought to the study of Hiroshima and Nagasaki victims, and how his demeanor set the tone for lively and cooperative exchanges throughout the meeting.

Existing methods had failed to detect an anticipated increase in mutations among the more than 12,000 children of Hirsohima–Nagasaki survivors (whose parents received an average 43 rad). Calculations showed that to measure a 30% increase in the mutation rate, roughly what would be expected from the average dose, one would have to examine 4.5×10^{10} bp in the children, and 4 to 5 times more in the parents (Delahanty, 1986). In fact, the DNA methods were at least an order of magnitude short of being able to detect the expected impact from atomic bomb exposure among survivors; they could only detect differences expected from radiation exposure well above the lethal dose (and hence not measurable). The question was whether there were new technical means

that would get around the problems. The answer was no, but the process of thinking about it forced many novel ideas to the surface.

George Church began to ruminate on the ideas that culminated in multiplex sequencing. He said later that discussions with Maynard Olson, Richard Myers, and others helped him crystallize his inchoate ideas. (David Smith recalled watching George Church disappear in a cloud of new-fallen powder one afternoon, and worrying about the future of DNA sequencing technology.)

Richard Myers showed work using RNase I to cut (and thus make detectable) single base pair mismatches; he and Leonard Lerman showed early data using gradients of denaturing agents embedded in electrophoresis gels as a way to detect heteroduplexes and mismatches. Myers credits his roommate for the conference, Maynard Olson, with clarifying his ideas and permitting him to expand the RNase I method to mismatches other than C–A mutations. In a trip report to the Office of Technology Assessment, Michael Gough characterized the Church and Myers presentations as technological wonders and called the two young scientists, then largely unknown, the "two biggest surprises" of the meeting (Gough, 1984).

Charles Cantor showed how his and David Schwartz's first pulsed-field gel electrophoresis method could separate megabase-sized DNA fragments, resolving individual yeast chromosomes and thus introducing an enormously powerful method to assess DNA structure on this scale. He also showed his and Cassandra Smith's first macrorestriction digest of the *Escherichia coli* genome, which suggested the tantalizing possibility of physically mapping entire genomes by combining restriction cleavage and pulsed-field gel electrophoresis.

Maynard Olson showed early results of attempting to construct a physical map of *Saccharomyces cerevisiae* using overlapping clones, and also showed good separation of megabase-sized DNA using a modification of the Schwartz-Cantor electrophoresis technique. Mendelsohn's DOE report noted that "while Olson's method would not presently be chosen for analyzing human mutation rates, his philosophy of paying careful attention to and investing in the quantitative, methodological details of DNA technology had a recurrent and important impact on the meeting" (Mendelsohn, 1985). Olson later brought the same core ideas to the National Research Council Committee on Mapping and Sequencing the Human Genome, where those ideas, combined with an expansion of goals to include genetic mapping, helped to forge a consensus that dedicated genome projects were scientifically worthwhile (National Research Council, 1988).

At Alta, Elbert Branscomb described the state of the art in using flow cytometry and immunofluorescence to detect altered protein products on the surface of red cells. Branscomb later became the computer modeler and one of the architects for the

Livermore cosmid map of chromosome 19, now under construction. Tom Caskey reviewed progress on understanding mutations in the HPRT locus, and Sherman Weissman reviewed data on the HLA locus. David Botstein, as always exuding volcanic enthusiasm peppered with sharp humor, speculated about pushing the restriction fragment length polymorphism (RFLP) techniques to their limits—perhaps enough to detect mutations in the range of 10^{-7} per base pair per generation. Unfortunately, this was still shy of what would be needed to detect mutations among the Hiroshima–Nagasaki survivors, unless an unrealistically massive effort were mounted. Ray White talked about applying RFLP methods to the Y chromosomes originating from a single Mormon progenitor of 1850 (who by now has thousands of male descendants) to examine changes in the part of the Y chromosome outside the pseudoautosomal region—a part of the genome where changes should accumulate.

Edwin Southern wound up the scientific session by addressing the gap between cytogenetic detection and molecular methods, and his presence was noted by more than one participant as a moderating influence on the intellectual pyrotechnics. Southern's discussion of measuring uv-induced mutations might be seen to presage the radiation hybrid mapping methods brought to fruition in 1988 by David Cox and Richard Myers, although the two approaches are quite independent in origin.

Michael Gough returned from Alta to Washington to work on the OTA report on detecting heritable mutations. The report had been requested by Congress in anticipation that controversies over Agent Orange, radiation exposure during atmospheric testing in the 1950s, and exposure to mutagenic chemicals might find their way to court, where a neutral assessment of the technical feasibility of detecting mutations would be essential. Gough directed preparation of *Technologies for Detecting Heritable Mutations in Human Beings* until he left OTA in 1985 (U.S. Congress, 1986).Several Alta participants served either as contractors or as advisory panel members for that study. Charles DeLisi, then newly appointed director of the Office of Health and Environmental Research at DOE, read a draft of this report in October 1985, and while reading it first had the idea for a dedicated human genome project (DeLisi, 1988). The Alta meeting is thus the bridge from DOE's traditional interest in detection of mutations to DeLisi's push for a Human Genome Initiative, and provides one of several historical links between genome projects and another massive technical undertaking of the 20th century—the Manhattan project.

Acknowledgements

Thanks go to the many Alta participants and others who reviewed drafts of this historical sketch—Elbert Branscomb, Charles Cantor, Charles DeLisi, Michael Gough, Mortimer Mendelsohn, Richard Myers, Maynard Olson, David Smith, and Ray White—and to those who helped provide background in interviews (see Ref. (4)).

Appendix B: "The Alta Summit"*

References

1. DELAHANTY, J., WHITE, R. L., AND MENDELSOHN, M. L. (1986). Approaches to determining mutation rates in human DNA. *Mutat. Res.* **167**: 215–232.

2. DELISI, C. (1988). The Human Genome Project. *Amer. Sci.* **76**: 488–493.

3. GOUGH, M. (1984). Notes from the DOE's Utah Meeting on DNA Methods for Measuring Human Mutation Rates, Trip Report to the Office of Technology Assessment, 21 December 1984.

4. Interviews with David Botstein, 22 August 1988; Charles Cantor, 19 August 1988; George Church, 14 November 1988; Charles DeLisi, 6 January 1987 and 7 October 1988; Maynard Olson, 28–30 April 1988; David Schwartz, 6 January 1987; and David Smith, 22 December 1988.

5. MENDELSOHN, M. L. (1985). Informal Report of a Meeting on DNA Methods for Measuring the Human Heritable Mutation Rate, Lawrence Livermore National Laboratory Report UCID-20315, January 1985.

6. National Research Council, Committee on Mapping and Sequencing the Human Genome (1988). "Mapping and Sequencing the Human Genome," National Academy Press, Washington, DC.

7. U.S. Congress, Office of Technology Assessment (1986). "Technologies for Detecting Heritable Mutations in Human Beings," OTA-H-298, U.S. Govt. Printing Office, Washington, DC.

*Reprinted with permission from Academic Press, Inc., *Genomics* **5**, 661–663 (October 1989).

Glossary

Appendix C

Appendix C: Glossary

Portions of the glossary text were taken directly or modified from definitions in the U.S. Congress Office of Technology Assessment document: *Mapping Our Genes—The Genome Projects: How Big, How Fast?* OTA-BA-373, Washington, D.C.: U.S. Government Printing Office, April 1988.

A word printed in a typeface different from that of the definition text is defined within the glossary.

Adenine (A): A nitrogenous base, one member of the base pair, A-T (adenine-thymine).

Alleles: Alternative forms of a genetic locus; a single allele for each locus is inherited separately from each parent (e.g., at a locus for eye color, a certain allele might result in brown eyes).

Amino acid: Any of a class of 20 molecules that are combined to form proteins in living things. The sequence of amino acids in a protein and hence protein function are determined by the genetic code.

Arrayed library: Arrayed libraries represent individual primary recombinant clones (hosted in phage, cosmid, YAC, or other vector) that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified based on the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications including screening for a specific gene or genomic region of interest as well as for physical mapping. Information gathered for individual clones from various genetic linkage and physical map analyses is entered into a relational database and used to construct physical and genetic linkage maps simultaneously; clone identifiers serve to interrelate the multilevel maps. Compare *library, genomic library.*

Autoradiography: A technique that uses X–ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel electrophoresis.

Autosome: A chromosome not involved in sex determination. The diploid human genome consists of 46 chromosomes, 22 pairs of autosomes, and 1 pair of sex chromosomes (the x and y chromosomes).

Bacteriophage: See *phage.*

Base pair (bp): Two nitrogenous bases (adenine and thymine or guanine and cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Base sequence analysis: A method, sometimes automated, for determining the sequence of nucleotide bases in DNA.

Blotting: See in situ *colony hybridization, Southern blotting.*

Blunt ends: On linear duplex DNA molecules, ends that are fully double-stranded and base-paired, without single-stranded tails.

Centimorgan (cM): A unit of measure of recombination frequency. One centimorgan is equal to a 1-percent chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million base pairs.

Centromere: A specialized chromosome region to which spindle fibers attach during cell division.

Chromosomes: The autoreplicating genetic structures of cells, containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome. Eukaryotic genomes are divided between a number of chromosomes in which the DNA is associated with many different kinds of proteins.

Clone bank: See *Genomic library.*

Cloning: The process of asexually producing a group of cells (clones), all genetically identical, from a single ancestor. In recombinant DNA technology, the use of DNA manipulation procedures to produce multiple copies of a single gene or segment of DNA is referred to as cloning DNA.

Cloning vector: DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artifical chromosomes; vectors are often recombinant molecules containing DNA sequences from several sources.

Code: See *genetic code.*

Cohesive (sticky) ends: On linear duplex DNA molecules, single-stranded ends that are complementary and can base-pair either with each other to form a circular molecule or with other linear DNAs having the same termini to form recombinant DNA molecules.

Colony hybridization: See *in situ colony hybridization.*

Complementary DNA (cDNA): DNA that is synthesized from a messenger RNA template; the single–strand form is often used as a probe in physical mapping.

Complementary sequences: Nucleic acid sequences that can form a double-stranded structure by formation of base pairs; the sequence complementary to G-T-A-C is C-A-T-G.

Appendix C: Glossary

Contigs: Groups of clones representing overlapping regions of a genome.

Cosmid: Artificially constructed cloning vector containing the *cos* gene of phage lambda. Cosmids can be packaged in lambda phage particles for infection into *E. coli*; this permits cloning of larger DNA fragments (up to 45 kb) than can be introduced into bacterial hosts in plasmid vectors.

Crossing over: The breaking during meiosis of one maternal and one paternal chromosome, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of alleles between chromosomes.

Cytosine (C): A nitrogenous base, one member of the base pair, G-C (guanine and cytosine).

C-value paradox: The lack of correlation between the amount of DNA in a haploid genome and the biological complexity of the organism. (C-value refers to haploid genome size.)

Determinism: The theory that for every action taken there are causal mechanisms such that no other action was possible.

Diploid: A full set of genetic material, consisting of paired chromosomes—one chromosome from each parental set. Most animal cells except the gametes have a diploid set of chromosomes. The diploid human genome has 46 chromosomes. Compare *haploid*.

DNA, deoxyribonucleic acid: The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus the sequence of each single strand can be deduced from that of its partner.

DNA probes: See *probes*.

DNA replication: The use of existing DNA as a template for the synthesis of new DNA strands. In humans and other eukaryotes, replication occurs in the cell nucleus.

DNA sequence: The relative order of base pairs, whether in a fragment of DNA, a gene, a chromosome, or an entire genome. See *base sequence analysis*.

Domain: A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function.

Double helix: The shape that two linear strands of DNA assume when bonded together.

Electrophoresis: A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Separation is based on these differences. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids.

Endonuclease: An enzyme that cleaves its nucleic acid substrate at internal sites in the nucleotide sequence.

Enzyme: A protein that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction.

Eukaryote: Cell or organism with membrane–bound, structurally discrete nucleus and other well–developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and blue–green algae. Compare *prokaryote*. See *chromosome*.

Exons: The protein–coding DNA sequences of a gene. Compare *introns*.

Exonuclease: An enzyme that cleaves nucleotides sequentially from free ends of a linear nucleic acid substrate.

Flow cytometry: Analysis of biological material by detection of the light-absorbing or fluorescing properties of cells or subcellular fractions (i.e., chromosomes) passing in a narrow stream through a laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, sort successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet.

Flow karyotyping: Use of flow cytometry to analyze and/or separate chromosomes on the basis of their DNA content.

Gamete: Mature male or female reproductive cell with a haploid set of chromosomes (23 for humans); that is, a sperm or ovum.

Gene: The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule). See *gene expression*.

Gene expression: The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs).

Appendix C: Glossary

Gene families: Groups of elosely related genes that make similar products.

Gene library: See *genomic library*.

Gene mapping: Determination of the relative positions of genes on a DNA molecule (chromosome or plasmid) and of the distance, in linkage units or physical units, between them.

Gene product: The biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease–causing alleles.

Genetic code: The sequence of nucleotides, coded in triplets along the mRNA, that determines the sequence of amino acids in protein synthesis. The DNA sequence of a gene can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence.

Genetic engineering technologies: See *recombinant DNA technologies*.

Genetics: The study of the patterns of inheritance of specific traits.

Genome: All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.

Genome projects: Research and technology development efforts aimed at mapping and sequencing some or all of the genome of human beings and other organisms.

Genomic library: A collection of clones made from a set of randomly generated overlapping DNA fragments representing tire entire genome of an organism. Compare *library, arrayed library*.

Guanine (G): A nitrogenous base, one member of the base pair, G-C (guanine and cytosine).

Haploid: A single set of chromosomes (half the full set of genetic material), present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. Compare *diploid*.

Heteroduplex: A double-stranded DNA molecule in which the two strands are not completely complementary in base sequence and hence are not completely base-paired.

Homeo box: A short stretch of nucleotides whose sequence is virtually identical in all the genes that contain it. It has been found in many organisms, from fruit flies to human beings. In the fruitfly, a homeo box appears to determine when particular groups of genes are expressed during development.

Human gene therapy: Insertion of normal DNA directly into cells to correct a genetic defect.

Human Genome Initiative: Collective name for several projects begun in 1986 by DOE to (1) create an ordered set of DNA segments from known chromosomal locations, (2) develop new computational methods for analyzing genetic map and DNA sequence data, and (3) develop new techniques and instruments for detecting and analyzing DNA. This initiative is now known as the Human Genome Program.

Hybridization: The process of joining two complementary strands of DNA, or of DNA and RNA together, to form a double–stranded molecule.

In situ colony hybridization: Use of a DNA or RNA probe to detect by in situ hybridization the presence of the complementary DNA sequence in cloned bacterial or cultured eukaryotic cells.

Informatics: The study of the application of computer and statistical techniques to the management of information. In genome projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence information, and to predict protein sequence and structure from DNA sequence data.

International technology transfer: Movement of inventions and technical know–how across national borders.

Introns: The DNA sequences interrupting the protein–coding sequences of a gene; these sequences are transcribed into RNA but are cut out of the message before it is translated into protein. Compare *exons*.

Karyotype: A photomicrograph of an individual's chromosomes arranged in a standard format showing the number, size, and shape of each chromosome type; used in low–resolution physical mapping to correlate gross chromosomal abnormalities with the characteristics of specific diseases.

Kilobase (kb): Unit of length for DNA fragments on physical maps (equal to the distance spanned by 1000 base pairs).

Library: An unordered collection of clones (i.e., cloned DNA from a particular organism), whose relationship can be established by physical mapping. Compare *genomic library, arrayed library*.

Linkage: The proximity of two or more markers (e.g., genes, RFLP markers) on a chromosome; the closer together the markers are, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together.

Appendix C: Glossary

Linkage map: A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans.

Locus (plural: loci): The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position. Some restrict use of *locus* to regions of DNA that are expressed. See *gene expression*.

Mapping: See *gene mapping, linkage map, physical map*.

Marker: An identifiable physical location on a chromosome (e.g., restriction enzyme cutting site, gene) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined. See *RFLP, restriction fragment length polymorphism*.

Meiosis: The process of two consecutive cell divisions in the diploid progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a haploid set of chromosomes.

Messenger RNA, mRNA: RNA that serves as a template for protein synthesis. See *genetic code*.

Multifactorial or multigenic disorders: See *polygenic disorders*.

Mutation: Any heritable change in DNA sequence. Compare *polymorphism*.

Nucleotide: A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule. See *DNA, base pair, RNA*.

Nucleus: The cellular organelle in eukaryotes that contains the genetic material.

Oncogene: A gene, one or more forms of which is associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth.

Phage: A virus for which the natural host is a bacterial cell.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest-resolution physical map is the banding patterns on the 24 different chromosomes; the highest-resolution map would be the complete nucleotide sequence of the chromosomes.

Plasmid: Autonomously replicating, extrachromosomal circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors.

Polygenic disorders: Genetic disorders resulting from the combined action of alleles of more than one gene (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles, thus the hereditary patterns are usually more complex than those of single-gene disorders. Compare *single-gene disorders*.

Polymerase, DNA or RNA: Enzymes that catalyze the synthesis of nucleic acids on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides.

Polymerase chain reaction (PCR): A method for amplifying a DNA sequence using the Klenow fragment of *E. coli* DNA polymerase I and two 20-base primers, one complementary to the (+)-strand at one end of the sequence to be amplified and the other complementary to the (-)-strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation and dissociation produce rapid and highly specific amplification of the desired sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample.

Polymorphism: Difference in DNA sequence among individuals. Genetic variations occurring in more than 1 percent of a population would be considered useful polymorphisms for genetic linkage analysis. Compare *mutation*.

Primer: Short preexisting polynucleotide chain to which new deoxyribonucleotides can be added by DNA polymerase.

Probe: Single-stranded DNA or RNA molecules of specific sequence, labeled either radioactively or immunologically, that are used to detect the complementary base sequence by hybridization.

Prokaryote: Cell or organism lacking membrane-bound, structurally discrete nucleus and subcellular compartments. Bacteria are prokaryotes. Compare *eukaryote*. See *chromosome*.

Promoter: A site on DNA to which RNA polymerase will bind and initiate transcription.

Protein: A large molecule composed of one or more chains of amino acids in a specific sequence; the sequence is determined by the sequence of nucleotides in

the gene coding for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs, and each protein has unique functions. Examples are hormones, enzymes, and antibodies.

Recombinant DNA technologies: Procedures used to join together DNA segments in a cell–free system (an environment outside of a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular chromosome.

Resolution: Degree of molecular detail on a physical map of DNA, ranging from low to high.

Restriction enzyme, endonuclease: A protein that recognizes specific, short nucleotide sequences and cuts DNA at those sites. There are over 400 such enzymes in bacteria that recognize over 100 different DNA sequences. See *restriction enzyme cutting site*.

Restriction enzyme cutting site: A specific nucleotide sequence of DNA at which a particular restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs), others much less frequently (e.g., every 10,000 base pairs).

RFLP, restriction fragment length polymorphism: Variation, between individuals, in DNA fragment sizes cut by specific restriction enzymes; polymorphic sequences that result in RFLPs are used as markers on both physical maps and genetic linkage maps. RFLPs are usually caused by mutation at a cutting site. See *marker*.

Ribosomal RNA, rRNA: A class of RNA found in the ribosomes of cells.

RNA, ribonucleic acid: A chemical found in the nucleus and cytoplasm of cells; it plays an important role in protein synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including messenger RNA, transfer RNA, ribosomal RNA, and other small RNAs, each serving a different purpose.

Sequence: The order of the nucleotides in a nucleic acid or order of amino acids in a protein.

Sequence-tagged sites (STSs): Short (200–500 base pairs) DNA sequences that have a single occurrence in the human genome and whose location and base sequence are known. Detectable by polymerase chain reaction, STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and could serve as landmarks on the developing physical map of the human genome.

Sex chromosomes: The X and Y chromosomes in human beings that determine the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a karyotype. Compare *autosomes*.

Shotgun method: Cloning of DNA fragments randomly generated from a genome. See *library, genomic library*.

Shuttle vectors: Cloning vectors that are capable of replicating in both prokaryotic and eukaryotic hosts.

Single-gene disorder: Hereditary disorder caused by a mutant allele of a single gene (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). Compare *polygenic disorders*.

Somatic cells: Any cell in the body except gametes and their precursors.

Southern blotting: Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific sequences by radiolabeled complementary probes.

Spheroplast: Yeast or bacterial cell from which most of the cell wall has been removed by enzymatic or chemical treatment.

Sticky ends: See *cohesive ends*.

Technology transfer: The process of moving scientific findings into the commercial sector for conversion to useful products.

Telomere: The ends of chromosomes. These specialized structures are involved in the replication and stability of linear DNA molecules. See *DNA replication*.

Thymine (T): A nitrogenous base, one member of the base pair, A-T (Adenine-Thymine).

Transcription: The synthesis of an RNA copy from a sequence of DNA (a gene); the first step in gene expression. Compare *translation*.

Transfer RNA, (tRNA): A class of RNA having structures with triplet nucleotide sequences that are complementary to the triplet nucleotide coding sequences of mRNA. The role of tRNAs in protein synthesis is to bond with amino acids and transfer them to the ribosomes, where proteins are assembled according to the genetic code carried by mRNA.

Transformation: A process by which the genetic information carried by an individual cell is altered by incorporation of exogenous DNA into its genome.

Appendix C: Glossary

Translation: The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. Compare *transcription.*

Uracil: A nitrogenous base normally found in RNA but not DNA; uracil is capable of forming a base pair with adenine.

Vector: See *cloning vector.*

Virus: A noncellular biological entity that can reproduce only within a host cell. Viruses consist of nucleic acid covered by protein; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus.

VLSI: Computer jargon: literally, "very large system integrated" (i.e., 10,000 to 100,000 transistors on a chip).

Yeast artificial chromosomes (YACs): Cloning vectors [containing centromere (CEN) and autonomous-replication sequences (ARS)] that are derived from yeast chromosomes, eukaryotic telomere sequences, and a number of biochemical marker genes.

Acronym List

AEC	Atomic Energy Commission
ANL*	Argonne National Laboratory, Argonne, IL
ATCC	American Type Culture Collection, Rockville, MD
BNL*	Brookhaven National Laboratory, Upton, NY
CEPH	Centre d'Étude du Polymorphisme Humain
DOE	Department of Energy
ERDA	Energy Research and Development Administration
FCCSET	Federal Coordinating Council on Science, Engineering and Technology
HERAC*	Health and Environmental Research Advisory Committee
HGCC*	Human Genome Coordinating Committee
HGMIS*	Human Genome Management Information System (ORNL)
HUGO	Human Genome Organisation (international)
JITF*†	Joint Informatics Task Force
LANL*	Los Alamos National Laboratory, Los Alamos, NM
LBL*	Lawrence Berkeley Laboratory, Berkeley, CA
LLNL*	Lawrence Livermore National Laboratory, Livermore, CA
NAS	National Academy of Sciences (U.S.)
NIH†	National Institutes of Health, Bethesda, MD
NLGLP*	National Laboratory Gene Library Project (LANL, LLNL)
NRC	National Research Council (NAS)
OHER*	Office of Health and Environmental Research
ORNL	Oak Ridge National Laboratory, Oak Ridge, TN
OSTP	Office of Scientific and Technology Policy (White House)
OTA	Office of Technology Assessment (U.S. Congress)
PAC†	Program Advisory Committee on the Human Genome (NIH)
PNL*	Pacific Northwest Laboratory, Richland, WA
SBIR	Small Business Innovative Research

* Denotes U.S. Department of Energy organizations.
† Denotes U.S. Department of Health and Human Services organizations.

UNITED STATES
DEPARTMENT OF ENERGY
WASHINGTON, D.C. 20545

ER-72